TILBURG UNIVERSITY

Tilburg University

Technologies on the stand

van den Berg, B.; Klaming, L.

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
van den Berg, B., & Klaming, L. (Eds.) (2011). *Technologies on the stand: Legal and ethical questions in neuroscience and robotics*. Wolf Legal Publishers (WLP).

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 04. May. 2023

Technologies on the stand:

Legal and ethical questions in neuroscience and robotics

TECHNOLOGIES ON THE STAND

LEGAL AND ETHICAL QUESTIONS IN NEUROSCIENCE AND ROBOTICS

edited by
BIBI VAN DEN BERG & LAURA KLAMING

a

Tilt
Tilburg Institute for Law, Technology, and Society

Technologies on the Stand
Legal and Ethical Questions in Neuroscience and Robotics
Bibi van den Berg, Laura Klaming (eds.)

ISBN: 978-90-5850-650-4

Published by Wolf Legal Publishers (WLP)
P.O. Box 31051
6503 CB Nijmegen
The Netherlands
Tel: +31 24 355 19 04
Fax: +31 84 837 67 00
E-Mail: info@wolfpublishers.nl
www.wolfpublishers.com

Cover design: Debbie Rovers & Ellen Knol

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic or mechanical, photocopying, recording or otherwise, without prior permission of the publisher. Whilst the authors, editors and publisher have tried to ensure the accuracy of this publication, the publisher, authors and editors cannot accept responsibility for any errors, omissions, statements, or mistakes and accept no responsibility for the use of the information presented in this work.

© WLP / Bibi van den Berg & Laura Klaming, 2011.

Technologies on the stand: Legal and ethical questions in neuroscience and robotics originates in a conference with the same title, which was held at Tilburg University on 11 and 12 April 2011. The editors wish to thank the following sponsors, for making this event possible:

BIRD & BIRD

CRIMINEE!

DE BRAUW BLACKSTONE WESTBROEK

HOUTHOFF BURUMA

surplus

PHILIPS
sense and simplicity

Springer

TILBURG UNIVERSITY
Law School

Van den Ende Productions©

Wetenschappelijk Onderzoek- en Documentatiecentrum
Ministerie van Veiligheid en Justitie

WOLF LEGAL PUBLISHERS

XS4ALL

Foreword

At present, neurotechnologies such as functional Magnetic Resonance Imaging (fMRI) and Deep Brain Stimulation are mainly used in the health sector for research, diagnosis and therapy. But neurotechnologies could also be used for human enhancement, for instance to improve cognitive functions or to morally enhance convicted offenders. Moreover, insights from neuroscience are increasingly used for legal purposes, for instance to determine a suspect's responsibility for his actions, or to distinguish truthful from deceptive statements. This raises the question whether neuroscience even has a contribution to make to (criminal) law at this point in time. Regardless of this concern, neuroscience has already entered the courtroom and influenced legal decisions. Using neurotechnologies for legal purposes obviously raises a number of important ethical and legal questions that require further discussion, most importantly regarding the admissibility of neurotechnologies in court.

Similarly, the application of robotics and autonomous technologies in various (social) situations, including the home, hospital environments, traffic and in war, raises a number of ethical and legal issues. These include questions such as: what are the ethical implications of applying robots in the health sector with regard to our ideas about human dignity and autonomy? What are the consequences of using robotics in war? And can we hold robots liable if they play an ever more important role in our daily lives? The increasing autonomy and intelligence of robotics technologies, moreover, raises questions regarding the moral and legal standing of such machines: should we implement ethics into robotic soldiers or robotic nannies, is this feasible, and if so, how should we go about designing moral machines?

Technologies on the stand: Legal and ethical questions in neuroscience and robotics is a textbook of papers that deal with diverse topics from the fields of law and neuroscience on the one hand and law, ethics and robotics on the other hand. The book is organised as follows: the **first part** deals with different topics from the field of law and neuroscience, ranging from criminal responsibility to the legal implications of using neuroscientific evidence, to human enhancement and its ethical and legal implications. The **second part** of the book deals with diverse topics from the field of law, ethics and robotics, and includes chapters on the morality of robots, the ethical and legal status of robots, and the regulation of behaviour through the design of robots.

Each part of the book is divided into three sections. We will now discuss each of these sections and the individual chapters contained in them. Section A of part I deals with criminal responsibility and the question to what extent neuroscience can be used to assess the responsibility of a suspect in a criminal trial. In **Chapter 1** Stephen Morse states that at present neuroscience does not have a large contribution to make to law generally, and that overclaims about the relevance of neuroscience to the law should be avoided. According to Stephen Morse, neuroscience is likely to make useful contributions to the law in the future by helping us to understand criminal behaviour better. **Chapter 2**, by Nicole Vincent, deals with restored mental capacities and the question whether direct brain interventions aimed at mental capacity restoration can help

us to assess the responsibility of someone who becomes mentally ill subsequent to committing their crime. In addition, the chapter addresses the question whether direct brain interventions aimed at mental capacity restoration help us to make a convicted offender more responsible.

Section B of part I deals with the legal issues raised by using neurotechnologies in the courtroom. In **Chapter 3**, Stefan Seiterle discusses the use of fMRI for lie detection as one of the core goals of criminal procedure. The main focus of this chapter is on the question whether, and under what circumstances, neuroscience-based lie detection would be admissible in criminal courts in Germany. **Chapter 4**, by Jan Christoph Bublitz deals with the ethical and legal issues of using neurotechnologies to change the minds of other people outside of a therapeutic context. Bublitz explains how neuroscience may change legal thinking about the protection of the mind. **Chapter 5**, by Laura Klaming, addresses one specific challenge of using neuroscience in the courtroom, i.e. the potentially overly persuasive influence of neuroscientific evidence on legal decision-making. More specifically, the importance of presentation mode in the discussion about the admissibility of neuroscientific evidence in court is emphasised. In **Chapter 6**, Tommaso Bruni discusses cross-cultural variability at the neural level and its consequences for the use of fMRI for the purpose of lie detection stressing that fMRI lie-detection may hinder the ascertainment of truth, if research does not take cross-cultural variability into account.

Section C of part one of *Technologies on the stand* deals with enhancement and the various ethical and legal questions that arise with regard to human enhancement. In **Chapter 7** Anna Pacholczyk discusses the use of neurotechnologies for the purpose of moral and social enhancement. Besides examining what we mean by moral enhancement and what is currently possible, she discusses the potential problems with morally enhancing interventions. **Chapter 8**, by Elizabeth Shaw, focuses on the possibility of employing neurotechnologies in the penal system to morally enhance offenders. Elizabeth Shaw argues against attempting to alter offenders' goals and values using neurotechnologies that wholly or largely circumvent the offender's rationality mainly for reasons of equality and moral dialogue. **Chapter 9**, by Bert-Jaap Koops and Ronald Leenes, deals with the possibility of using new technologies in order to improve our sight and vision and outlines a number of ethical and legal issues that may arise with this yet hypothetical form of human enhancement. In **Chapter 10** Pieter Bonte answers the question why we should be natural by presenting five arguments against the supposed duty to 'be natural' as grounds for outlawing human enhancement.

Part II of this book deals with law, ethics and robotics. Section A of part II addresses the foundations of roboethics. **Chapter 11**, by Wendell Wallach, focuses on ethics, law, and public policy in the development of robotics and neurotechnologies. Wallach argues that robotic technologies in combination with neurotechnologies and other emerging technologies will contribute to a transformation of human culture, which will pose important challenges that need to be addressed. In **Chapter 12**, Samir Chopra answers the question whether robots can be considered moral agents, focusing on the ascription of an appropriate set of beliefs and desires to a putative intentional entity. **Chapter 13**, by Steve Torrance, deals with the ethical and legal status of artificial agents. Specifically, the moral status of robots is linked to their consciousness. In **Chapter 14** David Jablonka addresses the problems that must be encountered and grappled with when discussing the moral responsibility of machines.

Section B of part II deals with ethics and the design of robots, and with the implementation of ethics or morality into robots. In **Chapter 15** Andreas Matthias analyses the concept of an ethical governor, which is supposed to effectively control and enforce the ethical use of lethal force by robots on the battlefield and which has had a great influence on the design of war robots. He argues that the concept of an ethical governor as favoured and already implemented by the military research community is misleading and does not address the moral problems it is supposed to solve. **Chapter 16**, by Aimee van Wynsberghe, outlines a framework for the ethical evaluation of care robots. Specifically, Aimee van Wynsberghe emphasises the importance of understanding the complexity of care practices, and the consequences this may have for designing care robots. In **Chapter 17** Joshua Lucas and Gary Comstock ask the question whether machines have *prima facie* duties by comparing two competing moral theories for the basis of algorithmic artificial ethical agents.

The final section of part II focuses on legal issues in robotics. **Chapter 18** deals with the legal responsibility of robots under Italian and European law. Chiara Boscarato discusses whether a robot should be considered an artefact or whether it should be compared to a person, for instance to a minor or a person with an unsound mind. In the **final chapter**, Bibi van den Berg argues that scholars in the field of Law & Technology ought to widen the scope of their research into techno-regulation, to include not only the *intentional* influencing of individuals through technological artefacts, but also more subtle, and implicit forms thereof. She discusses examples from various robotics domains to explain how this could work.

The editors wish to thank the following persons: first and foremost, the authors of the book, whose work has turned editing this volume into a real pleasure. We also wish to thank the reviewers for their time and effort to provide feedback on all of the papers. We thank Han Somsen and Anton Vedder, who, in their role as head of the Tilburg Institute of Law, Technology and Society (TILT) made it possible to organise the conference that was at the heart of this book. Thanks also to the members of the organising team who supported us in realising the conference and the book, Leonie de Jong, Femke Abousalama and Vivian Carter. We thank Debbie Rovers and Ellen Knol for the great job they did in designing promotional materials for the conference and the cover of this book. And last but not least, we thank our publisher, Simone Fennell, for a job well done.

Laura Klaming & Bibi van den Berg
Tilburg, the Netherlands, April 2011.

Contents

Author biographies

Pieter Bonte studied philosophy at Ghent University and law at the Free University of Brussels (VUB). In 2010 he started researching human enhancement at the Bioethics Institute Ghent. Tackling only the intrinsic arguments for and against human enhancement of one's own body (entailing also the possibility of parental enhancement of their offspring in the pre-birth stages of life) and thus excluding prudential arguments as well as questions of cultural and political interferences, he seeks to present a clearer picture of what may be wrong and/or right about human enhancement in itself. In the current, first stage of this project, he analyses the normativity of human nature, after which he will gauge the dramatically deepened responsibilities over ourselves and our offspring, to then conclude in 2013 with proposing a rational, useable conception of human dignity to handle the disruptive new liberties brought on by human enhancement technologies.

Chiara Boscarato graduated in 2009 in Law from the University of Pavia. Her final thesis, in Commercial Law, was on *The Responsibility of a Holding*. She is a trainee lawyer in Vigevano, Italy. Since February 2010 she has been working with the Interdepartmental Research Center ECLT – University of Pavia, Italy. On 1st December 2010 she became a Scholarship Fellow of the University of Pavia and ECLT Centre. Her research deals with the legal implications of the application of neuro techniques in the field of research and rehabilitation.

Tommaso Bruni is a PhD candidate in *Foundations of the Life Sciences and Their Ethical Consequences* at the University of Milan.

Jan Christoph Bublitz is a junior lecturer and researcher at the Faculty of Law, University of Hamburg. His research is at the intersection of law, moral philosophy and life sciences. His PhD thesis concerns the foundations and limits of a fundamental right of mental self-determination.

Samir Chopra is an Associate Professor of Philosophy at Brooklyn College and the Graduate Center of the City University of New York. Professor Chopra's interests include pragmatism, the philosophical foundations of artificial intelligence, the politics of technology, and legal theory. His latest work (co-authored with Laurence White), *A Legal Theory for Autonomous Artificial Agents* is forthcoming with the University of Michigan Press in April 2011. His previous work (co-authored with Scott Dexter), on the philosophical significance of free software, *Decoding Liberation: The Promise of Free and Open Source Software* was published by Routledge in 2007.

Gary L. Comstock is professor of philosophy at North Carolina State University where he conducts research on ethical questions in the biological sciences. He wrote the book, *Vexing nature: On the ethical case against agricultural biotechnology* and edited three other volumes.

David Jablonka is a PhD student at the University of Bristol (2010 to present) in the field of Philosophy of Law. He has an undergraduate degree in Law from the University of Kent (2006 -2009), and an LLM (Rearch Master) from the same university (2009 -2010). Jablonka is currently working on his PhD thesis with the working title *Roboethics and Legabotics – Can a machine be responsible for its actions?*

Laura Klaming holds a MSc degree in psychology from Maastricht University (2004) and a PhD (summa cum laude) from Bremen University (2008). At the Tilburg Institute for Law, Technology and Society (TILT), Laura's primary research interest lies in the area of law and neuroscience. Her current research at TILT concerns the possibility of applying neurotechnologies to various problems within the field of psychology and law, including improvement of eyewitness memory and the detection of deception, as well as the ethical and legal implications thereof. In addition, she is involved in research on the influence of neuroscientific evidence on legal decision-making.

Bert-Jaap Koops is Professor of Regulation & Technology at the Tilburg Institute for Law, Technology, and Society (TILT), the Netherlands. From 2005-2010, he was a member of *De Jonge Akademie*, a young-researcher branch of the Royal Netherlands Academy of Arts and Sciences. His research field is law & technology, in particular criminal-law issues such as cybercrime, cyber-investigation powers, and DNA forensics. He is also interested in other topics of technology regulation, including privacy, data protection, identity, digital constitutional rights, 'code as law', human enhancement, and regulation of bio- and nanotechnologies. From 2004-2009, he co-ordinated a VIDI research program on law, technology, and shifting power relations. Koops studied mathematics and general and comparative literature at Groningen University, and received his PhD in law at Tilburg University in 1999. He co-edited six books in English on ICT regulation, including *Starting Points for ICT Regulation* (2006) and *Dimensions of Technology Regulation* (2010).

Ronald Leenes is Professor of Regulation by Technology at the Tilburg Institute for Law, Technology, and Society (TILT), the Netherlands. His primary research interest is regulation by technology ('code' as law) in particular in the field of privacy and identity management, where he studies the role technology can play in protecting or enhancing privacy. He is also interested in other topics of technology regulation including ID fraud, digital constitutional rights, biometrics, robotics and Online Dispute Resolution. He has lead work packages in several EU FP6 and FP7 projects, including PRIME, PrimeLife, and ENDORSE. Leenes studied public administration at Twente University and received his PhD in law/public administration at Twente University in 1999. He co-edited six books in English on privacy, data protection, identity management, and ICT regulation, including *Privacy and Identity Management for Europe (PRIME)* (2011), and *Constitutional Rights and New Technologies, a comparative study* (2007).

Joshua Lucas is a student at North Carolina State University, where he is studying philosophy and

psychology. He is currently conducting research in the field of machine ethics.

Andreas Matthias studied philosophy, worked as a programmer and lecturer for programming languages for almost twenty years, before becoming a philosopher again. He has worked in Germany and Hong Kong, focusing on the moral aspects of computing technology, machine personhood, and the philosophical problems posed by artificial intelligence, technology, and the pursuit of happiness. He currently lives in Hong Kong and teaches at Lingnan University.

Stephen Morse J.D., PhD, is Ferdinand Wakeman Hubbell Professor of Law and Professor of Psychology and Law in Psychiatry at the University of Pennsylvania. Trained in both law and psychology at Harvard, Dr. Morse is an expert in criminal and mental health law whose work emphasises individual responsibility and the relation of the behavioural and neurosciences to responsibility and social control. Professor Morse was Co-Director of the MacArthur Foundation Law and Neuroscience Project and he co-directed the Project's Research Network on Criminal Responsibility and Prediction. He is co-editor with Adina Roskies of *A Primer on Law and Neuroscience* (forthcoming, Oxford University Press), and is currently working on a book, *Desert and Disease: Responsibility and Social Control*. Professor Morse is a founding director of the Neuroethics Society. Prior to joining the Penn faculty, he was the Orrin B. Evans Professor of Law, Psychiatry and the Behavioral Sciences at the University of Southern California.

Anna Pacholczyk studied for her BSc in Cognitive Science at the University of Westminster, and for her MA in Health Care Law and Ethics at Manchester University. She is currently a PhD student under the supervision of John Harris and Søren Holm. Her doctoral study considers ethics with regard to moral and social enhancement. Her principal research interests are in the ethics of enhancement and the social and ethical implications of the developments in neuroscience, including the ethics of use new technologies such as brain imaging, TMS and DBS as well as empirical investigations of morality and the consequences of this research for moral philosophy.

Stefan Seiterle is a research and teaching assistant of criminal law and criminal procedure law with the law faculty at the Europa-Universität Viadrina Frankfurt (Oder). He studied law at the universities of Konstanz, Amiens and Berlin (FU). In 2009, he obtained his doctorate degree (summa cum laude). 2009/10 he was a visiting fellow at the Zentrum für Interdisziplinäre Forschung (ZiF) in Bielefeld with the research group 'Challenges to the Image of Humanity and Human Dignity by New Developments in Medical Technology'. His research areas include medical criminal law, neurolaw, legal philosophy and bioethics.

Elizabeth Shaw holds an honours law degree (first class) from Aberdeen University (2008) and a masters degree in law (with distinction) from the same university (2010). She is undertaking a PhD at Edinburgh University on the topic of *Determinism, Criminal Responsibility and Punishment*. Her PhD attempts to develop a defensible non-retributive approach to the problem of criminal behaviour, which does not rely on a heavily contested notion of 'free will'. Currently she is working on the question of whether it is possible to explain what is problematic about 'manipulative' methods of dealing with criminal offenders (e.g. controlling

their behaviour via neurological interventions), without appealing to the ideas of 'free will' or 'retributive desert'. Her research is funded by The Arts and Humanities Research Council and by the Clark Foundation for Legal Education.

Steve Torrance is Visiting Senior Research Fellow at the Centre for Research in Cognitive Science (COGS) at the University of Sussex. He also teaches part-time at Goldsmiths College, London, and he is Emeritus Professor of Cognitive Science at Middlesex University. He writes on artificial ethics, on the feasibility and moral justifiability of creating artificial consciousness, and on the project to produce super-intelligent agents. He is currently a joint organiser of a workshop on Machine Consciousness this April in York, UK. He has edited a volume of the journal *AI and Society* on Ethics and Artificial Agents, and he has contributed a chapter to a forthcoming book on Machine Ethics.

Aimee van Wynsberghe is currently doing her PhD in Philosophy at the University of Twente, the Netherlands. During her undergraduate degree in Cell Biology at the University of Western Ontario, Canada, she was a research assistant at CSTAR (Canadian Surgical Technologies and Advanced Robotics) working on the Telesurgery project (long distance robotic surgery), which inspired her to continue working with robots. Following her studies in Science at UWO, she pursued a Masters in Applied Ethics at K.U. Leuven, Belgium and an Erasmus Mundus Masters in Bioethics. This gave her the opportunity to reflect on the philosophical issues pertaining to technology in healthcare, with a particular focus on robotics. Her current work focuses on the social and ethical implications of human-robot interactions, but specifically addresses the use of robots in the care of elderly persons by targeting issues of design.

Bibi van den Berg is post-doc researcher at the Tilburg Institute for Law, Technology and Society (TILT), at Tilburg University in the Netherlands. Her research areas are: (1) regulation and ethics in robotics, and (2) identity and privacy in online worlds. Van den Berg has a PhD in philosophy of technology, obtained from Erasmus University Rotterdam in the Netherlands in 2009.

Nicole Vincent obtained her PhD from the University of Adelaide in Australia in 2007 with a dissertation entitled *Responsibility, Compensation and Accident Law Reform*. She subsequently worked at Delft University of Technology in the Netherlands on a project entitled *The Brain and The Law*, which examined how neuroscience is relevant to legal responsibility. And since early 2011 she has been at Macquarie University in Sydney, Australia working on a project entitled *Reappraising the Capacitarian Foundation of Neurolaw*, which investigates whether doubts about the restoration and enhancement of responsibility challenge capacitarianism. She is also developing an Australasian Neurolaw Database.

Wendell Wallach is consultant, ethicist, and scholar at Yale University's Interdisciplinary Center for Bioethics. He chairs the Center's working research group on Technology and Ethics and is a member of other research groups on Animal Ethics, End of Life Issues, and Neuroethics. Wendell Wallach co-authored

(with Colin Allen) *Moral Machines: Teaching Robots Right From Wrong.* Formerly, he was a founder and the President of two computer consulting companies, Farpoint Solutions and Omnia Consulting Inc. Among the clients served by Mr. Wallach's companies were PepsiCo International, United Aircraft, and the State of Connecticut. Wallach is presently writing a book on the societal, ethical, legal, and public policy challenges posed by emerging and converging technologies. Another book in progress is *Cyber's Sowl*: *Self-Understanding in the Information Age,* which explores the ways in which cognitive science, new technologies, and introspective practices are altering our understanding of human decision making and ethics.

PART I: NEUROSCIENCE AND LAW

Section A: Neuroscience and responsibility

Chapter 1
NeuroLawExuberance: A plea for neuromodesty

Stephen J. Morse*
University of Pennsylvania
University of Pennsylvania Law School
✉ smorse@law.upenn.edu

Abstract This chapter suggests on conceptual and empirical grounds that at present neuroscience does not have a large contribution to make to criminal justice doctrine, adjudication and policy and to law generally despite the great advances in the science. Irrational exuberance and overclaims about the relevance should be avoided. It also explains why the new neuroscience does not present a radical challenge to current legal conceptions of agency and responsibility. Although present caution is warranted, the chapter concludes that in the near and intermediate term, as the science advances, neuroscience might well make helpful contributions to the law.

Introduction

In a 2002 editorial, the Economist warned, *"Genetics may yet threaten privacy, kill autonomy, make society homogeneous, and gut the concept of human nature. But neuroscience could do all those things first."* (2002, p. 77). The genome was fully sequenced in 2001 and there has not been one resulting major advance in therapeutic medicine since. Thus, even in its most natural domain, medicine, genetics has not had the far-reaching consequences that were envisioned. The same has been true for various other sciences that were predicted to revolutionize the law, including behavioural psychology, sociology, psychodynamic psychology, and others. I believe that this will also be true of neuroscience, which is simply the newest science on the block. Neuroscience is not going to do the terrible things the Economist fears, at least not for the foreseeable future. Neuroscience has many things to say, but not nearly as much as people would hope,

* Ferdinand Wakeman Hubbell Professor of Law & Professor of Psychology and Law in Psychiatry, University of Pennsylvania. This paper was first prepared for and presented at a conference on October 22, 2010 sponsored by the Mercer Law Review, 'Brain Sciences in the Courtroom'. It will be published in 62 Mercer Law Review (2011). It is published here in altered form with the kind permission of the editors of the Mercer Law Review. Jakob Elster and Michael Moore provided invaluable insights. As always, I thank my personal attorney, Jean Avnet Morse, for her sound, sober counsel and moral support.

especially in relation to law. At most, in the near to intermediate term, neuroscience may make modest contributions to legal policy and case adjudication. Nonetheless, there has been irrational exuberance about the potential contribution, an issue I addressed previously in an article addressing 'Brain Overclaim Syndrome' (Morse, 2006). I now wish to re-examine the case for caution.

In this chapter, I shall first make some remarks about the law's motivation and the motivation of some advocates to turn to science to solve the very hard normative problems that law addresses. Then I shall consider the law's psychology and its concept of the person and responsibility. I then consider how neuroscience might be related to law, which I call the issue of 'translation'. Next, I turn to various distractions that have bedeviled clear thinking about the relation of scientific, causal accounts of behaviour to responsibility. The chapter then examines the limits of neurolaw. The next section considers why neurolaw does not pose a genuinely radical challenge to the law's concepts of the person and responsibility. Nonetheless, the next section makes a case for cautious optimism about the contribution neuroscience may make to law in the near and intermediate term. A brief conclusion follows.

Science and law

Everyone understands that legal issues are normative, addressing how we should regulate our lives in a complex society. How do we live together? What are the duties we owe each other? When, for violation of those duties, is the State justified in imposing the most afflictive but sometimes justified exercises of state power, criminal blame and punishment?[1] When should we do this, to whom and how much?

Virtually every legal issue is contested – consider criminal responsibility, for example – and there is always room for debate about policy, doctrine and adjudication. In a fine, recent book, Professor Robin Feldman (2009) has argued that law lacks the courage forthrightly to address the difficult normative issues that it faces. It therefore adopts what Feldman terms an internalizing and an externalizing strategy for using science to try to avoid the difficulties. In the former, the law adopts scientific criteria as legal criteria. A futuristic example might be using neural criteria for criminal responsibility. In the latter, the law turns to scientific or clinical experts to make the decision. An example would be using forensic clinicians to decide whether a criminal defendant is competent to stand trial and then simply rubberstamping the clinician's opinion. Neither strategy is successful because each avoids facing the hard questions and they retard legal evolution and progress. Professor Feldman concludes, and I agree (Morse, 2011), that the law does not err by using science too little, as is commonly claimed. Rather, it errs by using it too much because the law is so insecure about its resources and capacities to do justice.

[1] In re Winship, 397 U.S. 358 (1970).

Here is my speculative interpretation of the motivation of enthusiasts for using neuroscience in criminal justice. Many hate the concept of retributive justice, thinking it is both prescientific and harsh. Their hope is that the new neuroscience will convince the law at last that determinism is true, that no offender is genuinely responsible, and that the only logical conclusion is that the law should adopt a consequentially-based prediction/prevention system of social control guided by the knowledge of the neuroscientist-kings who will finally have supplanted the Platonic philosopher-kings (Greene & Cohen, 2006, pp. 217-218). On a more modest level, many advocates think that neuroscience may not revolutionize criminal justice, but it will demonstrate that many more offenders should be excused and do not deserve the harsh punishments United States criminal justice imposes. Four decades ago, they would have been using psychodynamic psychology for the same purpose and more recently genetics has been employed similarly. The impulse is clear, however: jettison desert, or at least mitigate judgments of desert. As we shall see below, however, these advocates often adopt an untenable theory of mitigation or excuse that quickly collapses into the nihilistic conclusion that no one is really criminally responsible.

The law's psychology, concept of the person and responsibility

Criminal law presupposes a 'folk psychological' view of the person and behaviour. This psychological theory explains behaviour in part by mental states such as desires, beliefs, intentions, willings, and plans. Biological, other psychological and sociological variables also play a causal role, but folk psychology considers mental states fundamental to a full causal explanation and understanding of human action. Lawyers, philosophers and scientists argue about the definitions of mental states and theories of action, but that does not undermine the general claim that mental states are fundamental. Indeed, the arguments and evidence disputants use to convince others presuppose the folk psychological view of the person. Brains don't convince each other; people do. Folk psychology presupposes only that human action will at least be rationalisable by mental state explanations or that it will be responsive to reasons, including incentives, under the right conditions.

For example, the folk psychological explanation for why you are reading this chapter is, roughly, that you desire to understand the relation of neuroscience to criminal responsibility or to law generally, you believe that reading the chapter will help fulfil that desire, and thus you formed the intention to read it. This is a practical rather than a deductive syllogism.

Brief reflection should indicate that the law's psychology must be a folk psychological theory, a view of the person as a conscious (and potentially self-conscious) creature who forms and acts on intentions that are the product of the person's other mental states. We are the sort of creatures that can act for and respond to reasons. The law treats persons generally as intentional creatures and not simply as mechanistic forces of nature.

Law is primarily action-guiding and could not guide people directly and indirectly unless people could use rules as premises in their reasoning about how they should behave. Otherwise, law as an action-guiding system of rules would be useless, and perhaps incoherent. Legal rules are action-guiding primarily because they provide an agent with good moral or prudential reasons for forbearance or action. Human behaviour can

be modified by means other than influencing deliberation and human beings do not always deliberate before they act. Nonetheless, the law presupposes folk psychology, even when we most habitually follow the legal rules. Unless people are capable of understanding and then using legal rules to guide their conduct, law would be powerless to affect human behaviour.

The legal view of the person does not hold that people must always reason or consistently behave rationally according to some pre-ordained, normative notion of rationality. Rather the law's view is that people are capable of acting for reasons and are capable of minimal rationality according to predominantly conventional, socially-constructed standards. The type of rationality the law requires is the ordinary person's common sense view of rationality, not the technical notion that might be acceptable within the disciplines of economics, philosophy, psychology, computer science, and the like.

Virtually everything for which agents deserve to be praised, blamed, rewarded, or punished is the product of mental causation and, in principle, responsive to reason, including incentives. Machines may cause harm, but they cannot do wrong and they cannot violate expectations about how people ought to live together. Machines do not deserve praise, blame, reward, punishment, concern or respect because they exist or because of the results they cause. Only people, intentional agents with the potential to act, can violate expectations of what they owe each other and only people can do wrong.

Many scientists and some philosophers of mind and action consider folk psychology to be a primitive or pre-scientific view of human behaviour. For the foreseeable future, however, the law will be based on the folk psychological model of the person and behaviour described. Until and unless scientific discoveries convince us that our view of ourselves is radically wrong, the basic explanatory apparatus of folk psychology will remain central. It is vital that we not lose sight of this model lest we fall into confusion when various claims based on neuroscience are made. If any science is to have appropriate influence on current law and legal decision making, it must be relevant to and translated into the law's folk psychological framework, as shall be discussed in more detail below.

All of the law's doctrinal criteria for criminal responsibility are folk psychological. Begin with the definitional criteria, the 'elements' of crime. The 'voluntary' act requirement is defined, roughly, as an *intentional* bodily movement (or omission in cases in which the person has a duty to act) done in a reasonably integrated state of consciousness. Other than crimes of strict liability, all crimes also require a culpable further mental state, such as purpose, knowledge or recklessness. All affirmative defences of justification and excuse involve an inquiry into the person's mental state, such as the belief that self-defensive force was necessary or the lack of knowledge of right from wrong.

Our folk psychological concepts of criminal responsibility follow logically from the action-guiding nature of law itself, from its folk psychological concept of the person and action, and from the aim of achieving retributive justice, which holds that no one should be punished unless they deserve it and no more than they deserve. The general capacity for rationality is the primary condition for responsibility and the lack of that capacity is the primary condition for excusing a person. If human beings were not rational creatures who could understand the good reasons for action and were not capable of conforming to legal requirements through intentional action or forbearance, the law could not adequately guide action and it would not be just.

Legally responsible agents are therefore people who have the general capacity to grasp and be guided by good reason in particular legal contexts.[2]

In most cases of excuse, the agent who has done something wrong acts for a reason, but either is not capable of rationality generally or is incapable on the specific occasion in question. This explains, for example, why young children and some people with mental disorders are not held responsible. How much lack of rational capacity is necessary to find the agent not responsible is a moral, social, political, and ultimately legal issue. It is not a scientific, medical, psychological, or psychiatric issue.

Compulsion or coercion is also an excusing condition. Literal compulsion exists when the person's bodily movement is a pure mechanism that is not rationalisable by the agent's desires, beliefs and intentions. These cases defeat the requirement of a 'voluntary act'. For example, a tremor or spasm produced by a neurological disorder is not an action because it is not intentional and it therefore defeats the ascription of a voluntary act. Metaphorical compulsion exists when the agent acts intentionally, but in response to some hard choice imposed on the agent through no fault of his or her own. For example, if a miscreant holds a gun to an agent's head and threatens to kill her unless she kills another innocent person, it would be wrong to kill under these circumstances. Nevertheless, the law may decide as a normative matter to excuse the act of intentional killing because the agent was motivated by a threat so great that it would be supremely difficult for most citizens to resist. Cases involving internal compulsive states are more difficult to conceptualize because it is difficult to define 'loss of control' (Morse, 2002). The cases that most fit this category are 'disorders of desire', such as addictions and sexual disorders. The question is why these acting agents lack control but other people with strong desires do not? In any case, if the person frequently yields to his or her apparently very strong desires at great social, occupational, or legal cost to herself, the agent will often say that she could not help herself, that she was not in control, and that an excuse or mitigation was therefore warranted.

Lost in translation? Legal relevance and the need for translation

What in principle is the possible relation of neuroscience to law? We must begin with a distinction between internal relevance and external relevance. An internal contribution or critique accepts the general coherence and legitimacy of a set of legal doctrines, practices or institutions and attempts to explain or alter them. For example, an internal contribution of criminal responsibility may suggest the need for doctrinal reform, of, say, the insanity defence, but it would not suggest that the notion of criminal responsibility is itself incoherent or illegitimate. By contrast, an externally relevant critique suggests the doctrines, practices or institutions are incoherent, illegitimate or unjustified. Because a radical, external critique has little possibility

[2] I borrow the felicitious phrase, *"to grasp and be guided"* by reason from Wallace (1994).

of success at present, as I explain below, here I will make the simplifying assumption that the contributions of neuroscience will be internal and thus will need to be translated into the law's folk psychological concepts.

The law's criteria for responsibility and competence are essentially behavioural – acts and mental states. The criteria of neuroscience are mechanistic – neural structure and function. Is the apparent chasm between those two types of discourse bridgeable? This is a familiar question in the field of mental health law (Stone, 1984, pp. 95-96), but there is even greater dissonance in neurolaw. Psychiatry and psychology sometimes treat behaviour mechanistically, sometimes treat it folk psychologically, and sometimes blend the two. In many cases, the psychological sciences are quite close in approach to folk psychology. Neuroscience, in contrast, is purely mechanistic and eschews folk psychological concepts and discourse. Thus, the gap will be harder to bridge.

The brain does enable the mind, even if we do not know how this occurs. Therefore, facts we learn about brains in general or about a specific brain in principle could provide useful information about mental states and human capacities in general and in specific cases. Some believe that this conclusion is a category error (Bennett & Hacker, 2003; Pardo & Patterson, 2010). This is a plausible view and perhaps it is correct. If it is, then the whole subject of neurolaw is empty and there was no point to writing this chapter in the first place. Let us therefore bracket this pessimistic view and determine what follows from the more optimistic position that what we learn about the brain and nervous system can be potentially helpful to resolving questions of criminal responsibility if the findings are properly translated into the law's psychological framework.

The question is whether the new neuroscience is legally relevant because it makes a proposition about responsibility or competence more or less likely to be true. Any legal criterion must be established independently, and biological evidence must be translated into the criminal law's folk psychological criteria. That is, the expert must be able to explain precisely how the neuroevidence bears on whether the agent acted, formed a required mens rea, or met the criteria for an excusing condition. If the evidence is not directly relevant, the expert should be able to explain the chain of inference from the indirect evidence to the law's criteria. At present, as the part about the limits of neurolaw explains, few such data exist, but neuroscience is advancing so rapidly that such data may exist in the near or medium term. Moreover, the argument is conceptual and does not depend on any particular neuroscience findings.

Dangerous distractions concerning neuroscience and criminal responsibility and competence

This section of the article considers a number of related issues that are often thought to be relevant to criminal responsibility and competence but that are irrelevant or confusions and distractions: free will, causation as an excuse, causation as compulsion, prediction as an excuse, dualism, and the non-efficacy of mental states. Much of the legal exuberance about the contributions of neurolaw flow from these confusions and distractions, so it is important to correct them. But the legal exuberance also flows from unrealistic expectations about the scientific accomplishments of neuroscience. The next part of this article addresses the scientific exuberance.

Contrary to what many people believe and what judges and others sometimes say, free will is not a legal criterion that is part of any doctrine and it is not even foundational for criminal responsibility (Morse, 2007). Criminal law doctrines are fully consistent with the truth of determinism or universal causation that allegedly undermines the foundations of responsibility. Even if determinism is true, some people act and some people do not. Some people form prohibited mental states and some do not. Some people are legally insane or act under duress when they commit crimes, but most defendants are not legally insane or acting under duress. Moreover, these distinctions matter to moral and legal theories of responsibility and fairness that we have reason to endorse. Thus, law addresses problems genuinely related to responsibility, including consciousness, the formation of mental states such as intention and knowledge, the capacity for rationality, and compulsion, but it never addresses the presence or absence of free will.

When most people use the term free will or its lack in the context of legal responsibility, they are typically using this term loosely as a synonym for the conclusion that the defendant was or was not criminally responsible. They typically have reached this conclusion for reasons that do not involve free will, such as that the defendant was legally insane or acted under duress, but such usage of free will only perpetuates misunderstanding and confusion. Once the legal criteria for excuse have been met, for example—and none includes lack of free will as a criterion—the defendant will be excused without any reference whatsoever to free will as an independent ground for excuse.

There is a genuine metaphysical problem about free will, which is whether human beings have the capacity to act uncaused by anything other than themselves and whether this capacity is a necessary foundation for holding anyone legally or morally accountable for criminal conduct. Philosophers and others have debated these issues in various forms for millennia and there is no resolution in sight. Indeed, some people think the problem is not resolvable. This is a real philosophical issue, but, it is not a problem for the law, and neuroscience raises no new challenge to this conclusion. Solving the free will problem would have profound implications for responsibility doctrines and practices, such as blame and punishment, but, at present, having or lacking libertarian freedom is not a criterion of any civil or criminal law doctrine.

Neuroscience is simply the most recent mechanistic causal science that appears deterministically to explain behaviour. It thus joins social structural variables, behaviourism, genetics, and other scientific explanations that have also been deterministic explanations for behaviour. In principle, however, neuroscience adds nothing new, even if it is better, more persuasive science than some of its predecessors. No science, including neuroscience, can demonstrate that libertarian free will does or does not exist. As long as free will in the strong sense is not foundational for just blame and punishment and is not a criterion at the doctrinal level—which it is not—the truth of determinism or universal causation poses no threat to legal responsibility. Neuroscience may help shed light on folk psychological excusing conditions, such as automatism or insanity, for example, but the truth of determinism is not an excusing condition. The law will be fundamentally challenged only if neuroscience or any other science can conclusively demonstrate that the law's psychology is wrong and that we are not the type of creatures for whom mental states are causally effective. This is a different question from whether determinism undermines responsibility, however, and this article returns to it below.

A related confusion is that behaviour is excused if it is caused, but causation *per se* is not a legal or moral mitigating or excusing condition. I have termed this confusion *"the fundamental psycholegal error"* (Morse, 1994, pp. 1592-1594). At most, causal explanations can only provide evidence concerning whether a genuine excusing condition, such as lack of rational capacity, was present. For example, suppose a life history marked by poverty and abuse played a predisposing causal role in a defendant's criminal behaviour. Or suppose that an alleged new mental syndrome played a causal role in explaining criminal conduct. The claim is often made that such causes, which are not within the actor's capacity to control rationally, should be an excusing or mitigating position *per se*, but this claim is false.

All behaviour is the product of the necessary and sufficient causal conditions without which the behaviour would not have occurred, including brain causation, which is always part of the causal explanation for any behaviour. If causation were an excusing condition *per se*, then no one would be responsible for any behaviour. Some people welcome such a conclusion and believe that responsibility is impossible, but this is not the legal and moral world we inhabit. The law holds most adults responsible for most of their conduct and genuine excusing conditions are limited. Thus, unless the person's history or mental condition, for example, provides evidence of an existing excusing or mitigating condition, such as lack of rational capacity, there is no reason for excuse or mitigation.

Even a genuinely abnormal cause is not an excusing condition. For example, imagine a person with paranoid suspiciousness who constantly and hypervigilantly scans his environment for cues of an impending threat. Suppose our person with paranoia now spots a genuine threat that no normal person would have recognized and responds with proportionate defensive force. The paranoia played a causal role in explaining the behaviour, but no excusing condition obtained. If the paranoia produced a delusional belief that an attack was imminent, then a genuine excuse, legal insanity – an irrationality-based defence – might be appropriate.

In short, a neuroscientific causal explanation for criminal conduct, like any other type of causal explanation, does not *per se* mitigate or excuse. It provides only evidence that might help the law resolve whether a genuine excuse existed or it may in the future provide data that might be a guide to prophylactic or rehabilitative measures.

Compulsion is a genuine mitigating or excusing condition, but causation, including brain causation, is not the equivalent of compulsion. As we have seen, compulsion may be either literal or metaphorical and normative. It is crucial to recognize that most human action is not plausibly the result of either type of compulsion, but all human behaviour is caused by its necessary and sufficient causes, including brain causation. Even abnormal causes are not compelling. Suppose, for example, that a person with paedophilic urges has them weakly and is weakly sexed in general. If the person molested a child there would be no ground for a compulsion excuse. If causation were *per se* the equivalent of compulsion, all behaviour would be compelled and no one would be responsible. Once again, this is not a plausible account of the law's responsibility conditions. Causal information from neuroscience might help us resolve questions concerning whether legal compulsion existed or it might be a guide to prophylactic or rehabilitative measures when dealing with plausible legal compulsion. But causation is not *per se* compulsion.

Causal knowledge, whether from neuroscience or any other science, can enhance the accuracy of

behavioural predictions, but predictability is also not *per se* an excusing or mitigating condition, even if the predictability of the behaviour is perfect. To understand this, just consider how many things each of us does that are perfectly predictable for which there is no plausible excusing or mitigating condition. Even if the explanatory variables that enhance prediction are abnormal, excuse or mitigation is warranted only if a genuine excusing or mitigating condition is present. For example, recent research demonstrates that a history of childhood abuse coupled with a specific, genetically-produced enzyme abnormality that affects neurotransmitter levels vastly increase the risk that a person will behave antisocially as an adolescent or young adult (Caspi et al., 2002). A person is nine times more at risk if he has the MAOA deficiency and a childhood abuse history. Does that mean an offender with this gene by environment interaction is not responsible, or less responsible? No. The offender may not be fully responsible or responsible at all, but not because there is a causal explanation. What is the intermediary excusing or mitigating principle? Are these people, for instance, more impulsive? Are they lacking rationality? What is the actual excusal or mitigating condition? Again, causation is not compulsion and predictability is not an excuse. Just because an offender is caused to do something or is predictable does not mean the offender is compelled to do the crime charged or is otherwise not responsible. Brain causation, or any other kind of causation, does not mean we are automatons and not really acting agents at all.

Causal information may be of prophylactic or rehabilitative use for people affected, but no excuse or mitigation is applicable just because these variables make antisocial behaviour far more predictable. If the variables that enhance prediction also produce a genuine excusing or mitigating condition, then excuse or mitigation is justified for the latter reason and independent of the prediction.

Most informed people are not 'dualists' about the relation between the mind and the brain. That is, they no longer think that our minds (or souls) are independent of our brains (and bodies more generally) and can somehow exert a causal influence over our bodies. It may seem, therefore, as if law's emphasis on the importance of mental states as causing behaviour is based on a pre-scientific, outmoded form of dualism, but this is not the case. Although the brain enables the mind, we have no idea how this occurs and have no idea how action is possible (McHugh & Slavney, 1998, pp. 11-12). It is clear that, at the least, mental states are dependent upon or supervene on brain states, but neither neuroscience nor any other science has demonstrated that mental states play no independent and partial causal role. Indeed, the most likely explanation of complex human behaviour will be multi-field, multi-level, and will include mental states (Craver, 2007).

Despite our lack of understanding of the mind–brain-action relation, some scientists and philosophers question whether mental states have any causal effect, treating mental states as psychic appendixes that evolution has created but that have no genuine function. These claims are not strawpersons. They are seriously made by serious, thoughtful people (Greene & Cohen, 2006, pp. 217-218). As discussed below. If accepted, they would create a complete and revolutionary paradigm shift in the law of criminal responsibility and competence (and more widely). Thus, this claim is an external critique and must be understood as such. Moreover, given our current state of knowledge, there is little scientific or conceptual reason to accept it (Morse, 2011).

In conclusion, legal actors concerned with criminal law policy, doctrine and adjudication must always keep the folk psychological view present to their minds when considering claims or evidence from neuroscience and must always question how the science is legally relevant to the law's action and mental states criteria. The truth of determinism, causation and predictability do not in themselves answer any doctrinal or policy issue.

The limits of neurolaw

Most generally, the relation between brain, mind and action is one of the hardest problems in all science. We have no idea how the brain enables the mind or how action is possible (McHugh & Slavney, 1998). The brain-mind-action relation is a mystery. For example, we would like to know the difference between a neuromuscular spasm and intentionally moving one's arm in exactly the same way. The former is a purely mechanical motion, whereas the latter is an action, but we cannot explain the difference between the two. We know that a functioning brain is a necessary condition for having mental states and for acting. After all, if your brain is dead, you have no mental states, are not acting, and indeed are not doing much of anything at all. Still, we do not know how mental states and action are caused.

Despite the astonishing advances in neuroimaging and other neuroscientific methods, we still do not have sophisticated causal knowledge of how the brain works generally and we have little information that is legally relevant. This is unsurprising. The scientific problems are fearsomely difficult and only in the last decade have researchers begun to accumulate much data from functional magnetic resonance imaging (fMRI), which is the technology that has generated most of the legal interest. Moreover, virtually no studies have been performed to address specifically legal questions.

Before turning to the specific reasons for neuromodesty, a few preliminary points of general applicability must be addressed. The first and most important is to repeat the message of the prior section of this article. Causation by biological variables, including abnormal biological variables, does not *per se* create an excusing or mitigating condition. Any excusing condition must be established independently. The goal is always to translate the biological evidence into the criminal law's folk psychological criteria.

Assessing criminal responsibility involves a retrospective evaluation of the defendant's mental states at the time of the crime. No criminal wears a portable scanner or other neurodetection device that provides a measurement at the time of the crime. At least, not yet. Further, neuroscience is insufficiently developed to detect specific, legally-relevant mental content or to provide a sufficiently accurate diagnostic marker for even severe mental disorder (Frances, 2009). Nonetheless, certain aspects of neural structure and function that bear on legally relevant capacities, such as the capacity for rationality and control, may be temporally stable in general or in individual cases. If they are, neuroevidence may permit a reasonably valid retrospective inference about the defendant's rational and control capacities and their impact on criminal behaviour. This will of course depend on the existence of adequate science to do this. We now lack such science, but future research may remedy this.

Questions concerning competence or predictions of future behaviour are based on a subject's present condition. Thus, the retrospective problems besetting retrospective responsibility analysis do not apply to

such questions. The criteria for competence are functional. They ask whether the subject can perform some task, such as understanding the nature of a criminal proceeding or understanding a treatment option that is being offered, at a level the law considers normatively acceptable to warrant respecting the subject's choice and autonomy.

Now, let us begin consideration of the specific grounds for neuromodesty. At present, most neuroscience studies on human beings involve very small numbers of subjects, which makes establishing statistical significance difficult. Most of the studies have been done on college and university students, who are hardly a random sample of the population generally and of criminal offenders specifically. There is also a serious question of whether findings based on subjects' behaviour and brain activity in a scanner would apply to real world situations. Further, most studies average the neurodata over the subjects and the average finding may not accurately describe the brain structure or function of any actual subject in the study. Replications are few, which is especially important for law. Policy and adjudication should not be influenced by findings that are insufficiently established and replications of findings are crucial to our confidence in a result. Finally, the neuroscience of cognition and interpersonal behaviour is largely in its infancy and what is known is quite coarse-grained and correlational rather than fine-grained and causal (Miller, 2010).[3] What is being investigated is an association between a task in the scanner and brain activity. These studies do not demonstrate that the brain activity is either a necessary, sufficient or predisposing causal condition for the behavioural task that is being done in the scanner. Any language that suggests otherwise, such as claiming that some brain region is the neural substrate for the behaviour, is simply not justifiable. Moreover, activity in the same region may be associated with diametrically opposed behavioural phenomena, such as love and hate.

There are also technical and research design difficulties. It takes many mathematical transformations to get from the raw fMRI data to the images of the brain that are increasingly familiar. Explaining these transformations is beyond me, but I do understand that the likelihood that an investigator will find a statistically significant result depends on how the researcher sets the threshold for significance. There is dispute about this and the threshold levels are conventional. Change the threshold and the outcome will change. Now, I have been convinced by my neuroscience colleagues that many of such technical difficulties have been largely solved, but research design and potentially unjustified inferences from the studies are still an acute problem. It is extraordinarily difficult to control for all conceivable artifacts. Consequently, there are often problems of over-inference.

A major, potential problem for present and future collection and use of imaging evidence is whether an

[3] Miller (2010) also provides a cautious, thorough overview of the scientific and practical problems facing cognitive and social neuroscience.

uncooperative subject can invalidate a scan by the intentional use of countermeasures. This is not a problem if the subject either has a right not to be scanned, such as a 5th Amendment constitutional right in the United States not to be a witness against himself, or if the subject wishes to use neuroscience evidence. But if the subject can be scanned involuntarily or if the subject's purposes are served by invalidating a consensual scan, this is a difficulty. The first experimental study of this question has now been published and it discloses that in a laboratory lie-detection study, subjects could substantially undermine the accuracy of lie-detection by employing countermeasures (Ganis, Rosenfeld, Meixner, Kievit, & Schendan 2011).

Over time, however, these problems may ease as imaging and other techniques become less expensive and more accurate, as research designs become more sophisticated, and as the sophistication of the science increases generally. It is also an open question whether accurate inferences or predictions about individuals are possible using group data for a group that include the individual. This is a very controversial topic, but even if it is difficult or impossible now, it may become easier in the future.

Virtually all neuroscience studies of potential interest to the law involve some behaviour that has already been identified as of interest and the point of the study is to identify that behaviour's neural correlates. Neuroscientists do not go on general 'fishing' expeditions. There is usually some bit of behaviour, such as addiction, schizophrenia, or impulsivity, that they would like to understand better by investigating its neural correlates. To do this properly presupposes that the researchers have already identified and validated the behaviour under neuroscientific investigation. I call this the 'clear cut' problem. We typically get clear neuroscientific results only in cases in which the behavioural evidence was already clear.

On occasion, the neuroscience might suggest that the behaviour is not well-characterized or is neurally indistinguishable from other, seemingly different behaviour. In general, however, the existence of legally relevant behaviour will already be apparent. For example, some people are grossly out of touch with reality. If, as a result, they do not understand right from wrong, we excuse them because they lack such knowledge. We might learn a great deal about the neural correlates of such psychological abnormalities, but we already knew without neuroscientic data that these abnormalities existed and we had a firm view of their normative significance. In the future, however, we may learn more about the causal link between the brain and behaviour and studies may be devised that are more directly legally relevant. I suspect that we are unlikely to make substantial progress with neural assessment of mental content, but we are likely to learn more about capacities that will bear on excuse or mitigation.

The criteria for responsibility and competence criteria are behavioural; therefore, actions speak louder than images. This is a truism for all criminal responsibility and competence assessments. If the finding of any test or measurement of behaviour is contradicted by actual behavioural evidence, then we must believe the behavioural evidence because it is more direct and probative of the law's behavioural criteria. For example, if the person behaves rationally in a wide variety of circumstances, the agent is rational even if the brain appears structurally or functionally abnormal. And we confidently knew that some people were behaviourally abnormal, such as being psychotic, long before there were any psychological or neurological tests for such abnormalities. An analogy from physical medicine may be instructive. Suppose someone complains about back pain, a subjective symptom, and the question is whether the subject actually does have back pain. We

know that many people with abnormal spines do not experience back pain, and many people who complain of back pain have normal spines. If the person is claiming a disability and the spine looks dreadful, evidence that the person regularly exercises on a trampoline without difficulty indicates that there is no disability caused by back pain. If there is reason to suspect malingering, however, and there is not clear behavioural evidence of lack of pain, then a completely normal spine might be of use in deciding whether the claimant is malingering. Unless the correlation between the image and the legally relevant behaviour is very powerful, such evidence will be of limited help, however.

If actions speak louder than images, however, what room is there for using neuroevidence? Let us begin with cases in which the behavioural evidence is clear and permits an equally clear inference about the defendant's mental state. For example, lay people may not know the technical term to apply to people who are manifestly out of touch with reality, but they will readily recognize this unfortunate condition. No further tests of any sort will be necessary to prove this. In such cases, neuroevidence will be at most convergent and increase our confidence in what we already had confidently concluded. Whether it is worth collecting the neuroevidence will depend on how cost-benefit justified obtaining convergent evidence will be.

The most striking example of just such a case was the US Supreme Court's decision, *Roper v Simmons*,[4] which categorically excluded the death penalty for capital murders who killed when they were sixteen or seventeen years old because such killers did not deserve the death penalty. The *amicus* briefs were replete with neuroscience data showing that the brains of late adolescents are not fully biologically mature, and advocates used such data to suggest that the adolescent killers could not be fairly put to death. Now, we already knew from commonsense observation and rigorous behavioural studies that juveniles are on average less rational than adults. What did the neuroscientific evidence about the juvenile brain add? It was consistent with the undeniable behavioural data, and perhaps provided a partial causal explanation of the behavioural differences. The neuroscience data was therefore merely additive and only indirectly relevant and the Court did not cite it, except perhaps by implication.[5]

Whether adolescents are sufficiently less rational on average than adults to exclude them categorically from the death penalty is of course a normative legal question and not a scientific or psychological question. Advocates claimed, however, that the neuroscience confirmed that adolescents are insufficiently responsible

[4] *Roper v Simmons*, 543 US 551 (2005).

[5] The Court did refer generally to other science, but it was not clear if the neuroscience played a role. The Supreme Court did cite neuroscientific findings in Graham v. Florida, 560 U.S. (2010), which categorically excluded juveniles from life without the possibility of parole in non-homicide cases. The citation was general and I believe it was dictum. It was responding to an argument that no party had seriously made, which was that the science of adolescent development had changed significantly since *Roper* was decided.

to be executed, thus confusing the positive and the normative. The neuroscience evidence in no way independently confirms that adolescents are less responsible. If the behavioural differences between adolescents and adults were slight, it would not matter if their brains are quite different. Similarly, if the behavioural differences were sufficient for moral and constitutional differential treatment, then it would not matter if the brains were essentially indistinguishable.

If the behavioural data are not clear, then the potential contribution of neuroscience is large. Unfortunately, it is in just such cases that the neuroscience at present is not likely to be of much help. As noted, I term this the 'clear cut' problem. Recall that neuroscientific studies usually start with clear cases of well-characterized behaviour. In such cases, the neural markers might be quite sensitive to the already clearly identified behaviours precisely because the behaviour is so clear. Less clear behaviour is simply not studied or for less clear behaviour the overlap is greater between experimental and control subjects. Thus the neural markers of clear cases will provide little guidance to resolve behaviourally ambiguous cases of legally relevant behaviour. For example, suppose in an insanity defence case the question is whether the defendant suffers from a major mental disorder such as schizophrenia. In extreme cases, the behaviour will be clear and no neurodata will be necessary. Investigators have discovered various small but statistically significant differences in neural structure or function between people who are clearly suffering from schizophrenia and those who are not. Nonetheless, in a behaviourally unclear case, the overlap between data on the brains of people with schizophrenia and people without the disorder is so great that a scan is insufficiently sensitive to be used for diagnostic purposes.

Some people think that executive capacity, the congeries of cognitive and emotional capacities that help us plan and regulate our behaviour, is going to be the Holy Grail to help the law determine an offender's true culpability. After all, there is an attractive moral case that people with substantial lack of these capacities are less culpable, even if their conduct satisfied the prima facie case for the crime charged. Perhaps neuroscience can provide specific data previously unavailable to identify executive capacity differences more precisely. There are two problems, however. First, significant problems with executive capacity are readily apparent without testing and the criminal law simply will not adopt fine-grained culpability criteria. Second, the correlation between neuropsychological tests of executive capacity and actual real world behaviour is not terribly high (see Barkley & Murphy, 2010). Only a small fraction of the variance is accounted for, and the scanning studies will use the types of tasks the behavioural tests use. Consequently, we are far from able to use neuroscience accurately to assess non-obvious executive capacity differences that are valid in real world contexts.

Assessing the radical claim that we are not agents

This part of the chapter addresses the claim and hope alluded to earlier that neuroscience will cause a paradigm shift in criminal responsibility by demonstrating that we are 'merely victims of neuronal circumstances' or some similar claim that denies human agency, that holds that we are not the kinds of intentional creatures we think we are.

If our mental states play no role in our behaviour and are simply epiphenomenal, then traditional

notions of responsibility based on mental states and actions guided by mental states would be imperilled. But is the rich explanatory apparatus of intentionality simply a post-hoc rationalization the brains of hapless homo sapiens construct to explain what their brains have already done? Will the criminal justice system as we know it wither away as an outmoded relic of a prescientific and cruel age? If so, not only criminal law is in peril. What will be the fate of contracts, for example, when a biological machine that was formerly called a person claims that it should not be bound because it did not make a contract? The contract is also simply the outcome of various 'neuronal circumstances'.

Given how little we know about the brain-mind and brain-action connections, to claim based on neuroscience that we should radically change our picture of ourselves and our legal doctrines and practices is a form of neuroarrogance. Although I predict that we will see far more numerous attempts to introduce neuroevidence in the future, I have elsewhere argued that for conceptual and scientific reasons there is no reason at present to believe that we are not agents (Morse, 2008). It is possible that we are not agents, but the current science does not remotely demonstrate that this is true. The burden of persuasion is firmly on the proponents of the radical view.

What is more, the radical view entails no positive agenda. Suppose we were convinced by the mechanistic view that we are not intentional, rational agents after all. (Of course, the notion of being 'convinced' would be an illusion, too. Being convinced means that we are persuaded by evidence or argument, but a mechanism is not persuaded by anything. It is simply neurophysically transformed.) What should we do now? We know that it is an illusion to think that our deliberations and intentions have any causal efficacy in the world. We also know, however, that we experience sensations such as pleasure and pain and that we care about what happens to us and to the world. We cannot just sit quietly and wait for our brains to activate, for determinism to happen. We must, and will of course, deliberate and act.

If we still thought that the radical view were correct and that standard notions of genuine moral responsibility and desert were therefore impossible, we might nevertheless continue to believe that the law would not necessarily have to give up the concept of incentives. Indeed, Greene and Cohen concede that we would have to keep punishing people for practical purposes. Such an account would be consistent with 'black box' accounts of economic incentives that simply depend on the relation between inputs and outputs without considering the mind as a mediator between the two. For those who believe that a thoroughly naturalized account of human behaviour entails complete consequentialism, such a conclusion might not be unwelcome.

On the other hand, this view seems to entail the same internal contradiction just explored. What is the nature of the 'agent' that is discovering the laws governing how incentives shape behaviour? Could understanding and providing incentives via social norms and legal rules simply be epiphenomenal interpretations of what the brain has already done? How do 'we' 'decide' which behaviours to reward or punish? What role does 'reason' – a property of thoughts and agents, not a property of brains – play in this 'decision'?

If the truth of pure mechanism is a premise in deciding what to do, this premise yields no particular moral, legal or political conclusions. It will provide no guide as to how one should live or how one should

respond to the truth of reductive mechanism. Normativity depends on reason and thus the radical view is normatively inert. Neurons and neural networks do not have reasons; agents do. If reasons do not matter, then we have no reason to adopt any morals or politics, any legal rule, or to do anything at all.

Given what we know and have reason to do, the allegedly disappearing person remains fully visible and necessarily continues to act for good reasons, including the reasons currently to reject the radical view. We are not Pinocchios and our brains are not Giapettos pulling the strings.

The case for modest optimism

Despite my claim that we should be exceptionally cautious about the current contributions neuroscience can make to criminal law policy, doctrine, and adjudication, I am modestly optimistic about the possibility of near and intermediate term, limited but important contributions neuroscience can make to our ordinary, traditional, folk-psychological legal system. In other words, I think neuroscience may make a positive contribution although there has been no paradigm shift in thinking about the nature of the person and the criteria for criminal responsibility. The legal regime to which neuroscience will contribute will continue to take people seriously as people, as autonomous agents who may fairly be blamed and punished based on their mental states and actions.

In general, my hope is that over time there will be feedback between the folk psychological criteria and the neuroscientific data. Each might inform the other. Conceptual work on mental states might suggest new neuroscientific studies, for example, and the neuroscientific studies might help refine the folk psychological categories. The ultimate goal would be a reflective conceptual-empirical equilibrium. More specifically, there are four types of situations in which neuroscience may be of assistance that I will briefly address: data indicating that the folk wisdom underlying a legal rule is incorrect; data suggesting the need for new or reformed legal doctrine; evidence that helps adjudicate an individual case; and, data that help efficient adjudication or administration of criminal justice.

Many criminal law doctrines are based on bits of folk wisdom that may prove to be incorrect. If so, the doctrine should change. For example, it is commonly assumed that agents intend the natural and probable consequences of their actions. In many or most cases it seems that they do, but neuroscience may help in the future to demonstrate that this assumption is true far less frequently than we think. In that case, the rebuttable presumption used to help the prosecution prove intent should be softened or used with more caution.

Second, neuroscientific data may suggest the need for new or reformed legal doctrine. For example, control tests for legal insanity have been disfavoured for some decades because they are ill-understood and hard to assess. It is at present impossible to distinguish ‘can’t’ from ‘won’t’. Perhaps neuroscientific information will help to demonstrate and to prove the existence of control difficulties that are independent of cognitive incapacities. If so, then perhaps control tests are justified and can be rationally assessed after all. More generally, perhaps a larger percentage of offenders than we currently believe have such grave control difficulties that they deserve a generic mitigation claim that is not available in criminal law today. Neuroscience might help us discover that. If that were true, justice would be served by adopting a generic

mitigating doctrine. On the other hand, if it turns out that such difficulties are not so common, we could be more confident of the justice of current doctrine.

Third, neuroscience might provide data to help adjudicate individual cases. Consider the insanity defence again. There is often dispute about whether a defendant claiming legal insanity suffered from a mental disorder, about which disorder the defendant suffered from, and about how severe it was. At present, these questions must be resolved entirely behaviourally and there is often room for considerable disagreement about inferences drawn from the defendant's actions, including utterances. In the future, neuroscience might help resolve such questions if the clear cut problem difficulty can be solved. As mentioned previously, however, I doubt that, in the foreseeable future, neuroscience will be able to help identify the presence or absence of specific mens reas

Last, neuroscience might help us to implement current policy more efficiently. For example, the criminal justice system makes predictions about future dangerous behaviour for purposes of bail, sentencing, including capital sentencing, and parole. If we have already decided that it is justified to use dangerousness predictions to make such decisions, it is hard to imagine a rational argument for doing it less accurately if we were in fact able to do it more accurately. Behavioural prediction techniques already exist. The question is whether neuroscientific variables can provide value-added by increasing the accuracy of such predictions considering the cost of gathering such data. It is perfectly plausible that in the future they may and thus decisions will be more accurate and just.

Conclusion

At present, neuroscience has little to contribute to more just and accurate criminal law decision-making about policy, doctrine, and individual case adjudication. This was the conclusion I reached when I tentatively identified brain overclaim syndrome five years ago and it remains true today. In the future, however, as the philosophy of mind and action and neuroscience mutually mature and inform each other, neuroscience will help us understand criminal behaviour. No radical transformation of criminal justice is likely to occur. But neuroscience can inform criminal justice as long as it is relevant to law and translated into the law's folk psychological framework and criteria.

References

Barkley, R. A., & Murphy, K. R. (2010). Impairment in occupational functioning and adult ADHD: The predictive utility of executive function (EF) ratings versus EF tests. *Archives of Clinical Neuropsychology, 24*, 157-173.

Bennett, M. R., & Hacker, P. M. S. (2003). *Philosophical foundations of neuroscience*. Malden, MA: Blackwell Publishing.

Caspi, A., McClay, J, Moffit, T. E., Mill, J., Martin, J., Craig, I. W., Taylor, A., & Poulton, R. (2002). Role of genotype in the cycle of violence in maltreated children. *Science, 297*, 851-854.

Craver, C. F. (2007). *Explaining the brain*. New York: Oxford University Press.

Feldman, R. (2009). *The role of science in law*. New York: Oxford University Press.

Frances, A. (2009). Whither DSM-V? *The British Journal of Psychiatry, 195*, 391-392.

Ganis, G., Rosenfeld, J. P., Meixner, J., Kievit, R. A., & Schendan, H. E. (2011). Lying in the scanner: Covert countermeasures disrupt deception detection by functional magnetic resonance imaging. *NeuroImage, 55*, 312-319.

Greene, J., & Cohen, J. (2006). For the law, the new neuroscience changes nothing and everything. In S. Zeki & O. Goodenough (Eds.), *Law and the Brain* (207-227). New York: Oxford University Press.

McHugh, P., & Slavney, P. (1998). *Perspectives of psychiatry* (2nd Edition). Baltimore: The Johns Hopkins University Press.

Miller, G. A. (2010). Mistreating psychology in the decades of the brain. *Perspectives on Psychological Sciences, 5 (6)*, 716-743.

Morse, S. J. (1994). Culpability and control. *University of Pennsylvania Law Review, 142*, 1587-1660.

Morse, S. J. (2002). Uncontrollable urges and irrational people. *Virginia Law Review, 88 (5)*, 1025-1078.

Morse, S. J. (2006). Brain overclaim syndrome and criminal responsibility: A diagnostic note. Ohio State Journal of Criminal Law, 3 (2), 397-412.

Morse, S. J. (2007). The Non-Problem of Free Will in Forensic Psychiatry and Psychology. *Behavioral Sciences & the Law*, *25*, 203-220.

Morse, S. J. (2008). Determinism and the death of folk psychology: Two challenges to responsibility from neuroscience. *Minnesota Journal of Law, Science & Technology*, *9*(1), 1-36.

Morse, S. J. (2011). An accurate diagnosis, but is there a cure? *Hastings Science & Technology Law Journal, 3(1)*, 157.

Morse, S. J. (2011). Lost in translation: An essay on law and neuroscience. In M. Freeman (Ed.), *Law and Neuroscience* (529-562). New York: Oxford University Press.

Pardo, M., & Patterson, D. (2010). Philosophical foundations of neuroscience. *University of Illinois Law Review, 4,* 1211-1250.

Stone, A. A. (1984). *Law, psychiatry, and morality*. American Psychiatric Press.

The Economist (2002). The ethics of brain science: Open your mind. *The Economist, 25,* 77-79. Retrieved

from http://www.economist.com/node/1143317

Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Cambridge, MA: Harvard University Press.

Chapter 2
Capacitarianism, responsibility and restored mental capacities

Nicole Vincent
Macquarie University
Department of Philosophy
✉ nicole.vincent@mq.edu.au

Abstract The capacitarian idea that responsibility tracks mental capacity underlies much of our thinking about responsibility. For instance, mental capacity assessments inform whether someone is a fully responsible person, what responsibilities they can be expected to observe, their degree of responsibility for what they did, and whether they can be expected to take responsibility and be held responsible in the sense of standing trial, being answerable, paying compensation and being punished. But what happens when mental capacity is restored through direct brain interventions? Specifically, can direct brain interventions aimed at mental capacity restoration help us to assess the responsibility of someone who becomes mentally ill subsequent to committing their crime or to hold them responsible, to expect them to take responsibility for what they did, to make them fully responsible and maybe even less irresponsible? I will argue that initially capacitarianism seems to strike difficulties in cases that involve direct brain interventions of this sort, or put another way, that responsibility does not seem to track restored mental capacities. However, I will also argue that most of these difficulties can be overcome once we take into account some of the other things that responsibility also hinges upon. In particular, I will argue that historical and normative considerations can explain why responsibility does not seem to track restored mental capacities, and thus why this is not something that undermines capacitarianism.

Keywords capacitarianism, responsibility, direct brain interventions, therapy, justice

Capacitarianism underlies much of our thinking about responsibility

The 'capacitarian' idea that responsibility co-varies with or 'tracks' mental capacity underlies much lay, legal and philosophical thinking about responsibility. In *lay* contexts responsibility is often thought to require such things as the ability to perceive the world without delusion, to think clearly and rationally, to guide our actions by the light of our judgments, and to resist acting on mere impulse. This is, for instance, why children, the senile, and the mentally ill are thought to be less than fully responsible for what they do (i.e. because they lack the right kind and/or degree of mental capacity), why children can acquire more and/or greater responsibilities as they grow up (i.e. because their mental capacities develop as they mature), and how responsibility is reinstated on recovery from mental illness (i.e. because the needed mental capacities are recovered).

Identical sentiments are expressed in the *legal* context by the idea that responsibility requires certain 'cognitive' and 'volitional' mental capacities.[6] For instance, Hart (1968) suggests that *"[t]he capacities in question are those of understanding, reasoning, and control of conduct: the ability to understand what conduct legal rules or morality require, to deliberate and reach decisions concerning these requirements, and to conform to decisions when made"* (Hart, 1968, p. 227). As I have argued elsewhere (Vincent, 2011b), the capacitarian approach provides an effective framework for understanding how legal defences such as insanity and automatism serve to diminish an accused person's responsibility for something that they did at the guilt determination stage of a trial. Namely, they do this by highlighting deficits in the mental capacities which are required for full responsibility, and thus by undermining either the claim that the defendant is causally responsible for what they did (by challenging the claim that their behaviour should even be viewed as an instance of their action and thus as something which can be attributed to them) or that they violated their role responsibilities in acting as they did (by challenging the claim that the defendant acted contrary to how they ought to have acted and is thus blameworthy for acting like that). Indeed, many recent attempts to introduce neuroscience into courtrooms are also best understood in precisely this way – i.e. neuroscientific techniques are used to help make assessments of mental capacity which in turn inform assessments of responsibility (Vincent, 2010, 2011).

Mental capacity assessments also play a central role in the legal practice of holding defendants responsible (either in the sense of bringing them to trial where they must account for what they did or of punishing them for it) and of expecting them to take responsibility for what they did (in the sense of accounting for their actions during the trial, of acknowledging that they did it and are maybe even blameworthy, and perhaps that they might even need to do something like paying compensation or submitting themselves to punishment to make up for it). For instance, in *Connecticut v. Kenneth Curtis* (1999) a defendant who shot himself in the head after first shooting and killing his estranged girlfriend was initially found to be incompetent to stand trial, and thus he was initially neither tried nor punished for this crime. However, years later he made such an impressive recovery that he even enrolled in college, and upon subsequent re-assessment he was found competent to stand trial where he pled guilty to manslaughter and received a 20 year prison sentence. In *Ford v. Wainright* (1986), the court held that it was impermissible to punish defendants who are insane and thus whose responsibility-relevant mental capacities were compromised (though who were not insane at the time when they committed the crime) because this would serve no penological goals. And similar arguments were also cited in *Atkins v. Virginia* (2002) where the

[6] The word 'responsibility' refers to a range of different though related concepts and practices within the law. For a detailed discussion of the plural senses of 'responsibility' used in legal contexts, and why these are genuine responsibility concepts, see Vincent (2010, 2011b).

court held that it was impermissible to punish mentally retarded defendants. The idea that responsibility tracks mental capacity clearly plays a key role in the legal practices of holding people responsible and of expecting them to take responsibility.

Furthermore, consider the role that mental capacity assessments play in parole boards' investigations. One question here is whether the person under assessment has reformed sufficiently such that it is now safe to release them back into society – i.e. whether we have sufficient evidence that they are now a responsible individual who can be trusted to stay out of trouble. And another question concerns whether the person in question has sufficient mental capacities to be a fully responsible person in the sense that if they were released from prison, they would be sufficiently mentally competent to fend for themselves and should the need arise to be held responsible for anything else that they might do.

Finally, the idea that responsibility tracks mental capacity – and again, in particular, cognitive and volitional mental capacities – is also central to much *philosophical* thinking about this topic. For instance, Aristotle wrote that *"feelings and actions…that are involuntary receive pardon"* and that *"[a]ctions are regarded as involuntary when they are performed under compulsion or through ignorance"* (Aristotle, 1976, p. 111, emphasis added). Since Aristotle, this view has also been endorsed by numerous other philosophers (e.g. Dennett, 1984; Glannon, 2002; Sher, 2009; Wallace, 1994).

Furthermore, capacitarianism is also at the centre of Fischer's and Ravizza's (1998) compatibilist defence of responsibility from the alleged threat of determinism. Fischer and Ravizza (1998) contend that the reason why we retain the intuition that agents in Frankfurt style counter examples (Frankfurt, 1969, p. 835-836) are still responsible for what they did even though they could not have done otherwise is because they possessed 'guidance control' (Fischer & Ravizza, 1998). On their account, a person has guidance control over their actions when those actions issue from their own[7] *"moderately reasons-responsive mechanisms"*, or when via *"tracing" we establish that they are responsible for the fact that their mechanism was not moderately reasons-responsive*" (Fischer & Ravizaa, 1998, p. 49-51). Mechanisms are moderately reasons-responsive when they are *"regularly receptive"* and *"weakly reactive"* to reasons. And a mechanism is weakly reactive to reasons if it reacts to those reasons in at least a small set of possible worlds, even though it may not react to those reasons in this actual world (Fischer & Ravizza, 1998).

Put in simpler terms, on Fischer and Ravizza's (1998) account a person has guidance control and is thus responsible for what they do when their actions issue from mechanisms in virtue of which they possess the cognitive and volitional mental capacities which are required for moral agency. *Reasons-responsiveness* is a feature of certain brain mechanisms in virtue of which one can comprehend what one ought to do/not do, and be moved appropriately to action on the basis of that comprehension. And the *moderateness* component

[7] This refers to Fischer and Ravizza's ownership condition (1998).

of moderate reasons responsiveness captures the idea of these brain mechanisms having varying degrees of *capacity* to do those things, rather than being infallible clockwork-like mechanisms that function in a suspiciously regular and thus worryingly unfree fashion.

Capacitarianism and restored mental capacities

The previous section demonstrated the pervasive role that capacitarianism plays in our thinking about responsibility. I take this pervasiveness to indicate the strength of our commitment to capacitarian thinking – i.e. the strength of our conviction that responsibility really does track mental capacity.

In the examples cited above, what tracks mental capacity are *assessments* and *impositions* of responsibility. That is, mental capacity assessments inform assessments of whether someone should be considered a fully responsible person, assessments of the degree of their responsibility for something that they did, and assessments of whether they can be held responsible for what they have done in the sense of standing trial and being answerable. Similarly, mental capacity assessments also inform judgments about what sorts of responsibilities it is reasonable to impose onto people, and whether we can hold them responsible by imposing punishment and liability onto them. However, direct brain interventions are increasingly holding out the promise of making it possible to restore people's mental capacities, and when this is combined with the capacitarian idea that responsibility tracks mental capacity, two enticing legal prospects emerge – one for traditional *justice*-based aims, and another for the aim of *therapeutic* justice – which suggest that responsibility (in the various senses mentioned above) also tracks *restored* mental capacities.

Firstly, consider the way in which mental capacity restoration is already being used to bring criminals to *justice*. Criminals who develop mental disorders after committing crimes may lack competence to stand trial or to be punished, and this prevents them from taking responsibility and us from trying them to assess their responsibility and subsequently from holding them responsible for what they did. However, such mental disorders can nowadays sometimes be treated with medications or other medical procedures (Lekovic, 2008). For instance, defendants are sometimes involuntarily treated with antipsychotic drugs to restore their capacity to stand trial. In *Riggins v. Nevada* (1992), the court held that it was generally permissible to administer antipsychotic drugs that make defendants competent to stand trial as long as this is medically appropriate and the least intrusive means of making them competent.[8] In *Sell v. United States* (2003) the United States Supreme Court clarified this position by ruling that it was permissible to forcibly medicate a

[8] Notably, the court ruled that in this particular case Riggins had a legitimate reason to object to being medicated — namely, to support his insanity defense, Riggins wanted the jury to witness his condition first-hand without it being masked by the medications.

defendant for the *sole* purpose of making them competent to stand trial, as long as the treatment would most likely be effective, no medically better alternatives are available, and important state interests such as bringing criminals to justice are at stake. Similarly, defendants are sometimes also treated with antipsychotic drugs to restore their competence to be punished – for instance, in *Singleton v. Norris* (2003) antipsychotic medications were forcibly administered to a condemned defendant to make him competent for execution by lethal injection (Latzer, 2003). But even if we do not endorse capital punishment,[9] we might still endorse the more moderate position that inmates who develop mental disorders while serving a prison sentence should be treated for those disorders, and not only because prisoners retain a right to receive adequate medical care while incarcerated, but also to ensure that the retributive function of incarceration (one of the aforementioned penological goals cited in *Ford v. Wainright* (1986) above) is performed. My point is that such direct brain interventions seem capable of restoring people's ability to take responsibility, and our ability to assess their degree of responsibility for what they did and to subsequently hold them responsible for it – i.e. that responsibility in these different senses also seems to track restored mental capacities.

Secondly, consider some actual and proposed *therapeutic* uses of direct brain interventions. Some authors write that at least a portion of criminal behaviour may be caused by mental disease or disorder (Sapolsky, 2004; Tancredi, 2005), and so they suggest that such criminals should be provided with medical care and treatment rather than being punished. Dawkins (2006) argues that it makes just as little sense to punish criminals as it does for Basil Fawlty (the fictional character from the comedy 'Fawlty Towers') to threaten and then to flog his broken car when it fails to start. On Dawkins' account, instead of punishing criminals, we should figure out what's wrong with them – i.e. what brain disorder makes them act like that, for instance drug addiction – and then fix it by administering the appropriate treatments. Put in terms of the language of responsibility, what Dawkins is suggesting is that rather than holding irresponsible and non-responsible people responsible for what they did, we should instead treat their disorders and make them responsible.[10] It is for precisely these sorts of reasons that repeat sex offenders are sometimes given drugs like cyproterone acetate, a powerful antiandrogen that helps them to regain self-control by reducing their sex drive (Bradford & Pawlak, 1993). And this is also the thinking behind trials currently under way in Australia in which convicted violent criminals at treated with selective serotonin reuptake inhibitors (SSRIs), which have been observed to help people control their outbursts of anger – i.e. the aim is to restore their capacity for

[9] For the record, I do not endorse capital punishment, and nothing that I say in this paper should be taken to imply otherwise.

[10] Note however that Dawkins would probably not endorse putting his point into the language of responsibility since he explicitly argues that responsibility does not make sense in a fully mechanistic universe in which something like determinism is at work.

self-control (Macey, 2010).[11] Finally, Chinese and Russian surgeons have, rather controversially, reported success in treating drug addiction through ablation of the nucleus accumbens and cingulate gyrus, also arguably restoring the addicted individuals' capacity for self-control (Lekovic, 2008) – a technique which, if not so controversial, could perhaps be used as a treatment for those who commit drug-related crimes. Proponents of this therapeutic approach view it as a more compassionate and effective response to crime than just punishing criminals, because it allegedly targets the causes of crime – i.e. it aims to restore people's status as responsible individuals in both of the relevant senses – and so these examples also suggest that responsibility tracks restored mental capacities.

Problems with the idea that responsibility can be restored

The previous section argued that direct brain interventions which restore mental capacities could at least hypothetically help the law to achieve justice in the traditional and therapeutic senses. Or, put another way, it argued that direct brain interventions might help us to assess the responsibility of people who develop certain mental illnesses after committing their crime or to hold them responsible, to put us in a position where we can justifiably expect them to take responsibility for what they did, and even to make them fully responsible or less irresponsible – i.e. that responsibility in its various senses tracks restored mental capacities.

However, there are problems with these claims. Some of these are moral problems such as: practical ethics concerns about the risks and side-effects associated with any kind of treatments that involve direct brain interventions (Chatterjee, 2007); that it might be unethical, or at least violate codes of medical ethics, for members of the medical profession who are meant to save lives to treat people for the sole purpose of making them competent to be punished and maybe even executed (Eisenberg, 2004; Latzer, 2003); that people have a right to cognitive liberty (Tovino, 2007) – or, as Martha Farah puts it, *"the freedom to think one's own thoughts and have one's own personality"* (Farah, 2002, p. 1126) – which might be infringed even by voluntary let alone involuntary treatments of this sort (Bomann-Larsen, 2011); and that respect for human dignity is incompatible with modifying people to make them (in our view) better (Duff, 2005). However, rather than investigating these claims about what we *ought* and ought not to do, in what follows I wish to highlight difficulties with the underlying assumption that direct brain interventions are something that *could* even help us to achieve justice in traditional and therapeutic senses.

[11] Note though that in this trial the restoration of these convicted violent criminals' capacity for self-control, and thus of their responsibility, does not replace the traditional aims of justice since they are given no concession on their sentences.

Mental capacity restoration and justice

In regards to bringing criminals to justice, it may be that the sorts of medications that are typically administered to defendants to make them competent to stand trial or to be punished simply cannot restore their mental capacities, but that they either only make it *seem* as if the treated people again possess those mental capacities, or that they implant in them *artificial* mental capacities which are poor substitutes for their own mental capacities. For instance, Horstman has argued that:

> Artificial sanity is not a substitute for true sanity. Drugs inherently have only a temporary effect on people, usually just as long as the drug is in the blood stream. Additionally, the temporary nature of chemical competence renders difficult a reliable assessment of the sufficiency of an inmate's competence for execution. Since an insane inmate cannot truly be made permanently sane, it follows that the inmate cannot be made competent to be executed. (Horstman, 2002, pp. 846-847).

If what such treatments would do is more like either of these two things – i.e. forcing people to wear a *mask of sanity* or the creation of an *artificial sanity* – than genuine restoration of the treated person's original mental economy, then such techniques might be incapable of helping us to achieve the traditional aims of justice. In the former case, we would merely be making ourselves feel better about trying or punishing a person who in fact remains fundamentally mentally ill, and in the latter case the person who would be tried and/or punished would not even be the same as the person who committed the crime (Greely, 2008).

In a paper devoted to the topic of the 'medicate to execute controversy', Latzer (2003) considers the objection that *"[m]edication cannot achieve true competency for execution"* (2003, pp. 9-10). Although he ultimately argues that this objection is unconvincing, I believe that there are problems with his argument which help to highlight why we might doubt that direct brain interventions can achieve justice in the traditional sense, and since a portion of what Latzer (2003) says also applies to the use of other direct brain interventions (not just pharmaceuticals)[12] and to the aim of restoring competence to stand trial (not just competence to be punished), it is therefore worthwhile to consider his position in detail.

In regards to doubts about the effectiveness of such techniques (i.e. the 'mask of sanity' objection), Latzer (2003) acknowledges that *"[i]n a case in which the medication is ineffectual, the policy implications are clear: forcible (sic) medication would be pointless"*, however he maintains that such people *"are a distinct minority, they drive the exception, not the rule. For the majority, medication that controls (though it does not*

[12] I treat psychopharmaceuticals as an instance of direct brain interventions because their chemical method of action bypasses our conscious reasoning and control capacities and instead it targets the brain mechanisms that implement those capacities, albeit *via* our veins (in the case of injectable pharmaceuticals) or *via* our stomachs (in the case of pills).

cure) the disquieting effects of psychosis can produce competence for execution" (Latzer, 2003, p. 10). This leaves him with the 'artificial sanity' objection, which he dispatches as follows:

> [I]f the effects of the disorder can be controlled, and for as long as they can be controlled, then the inmate is competent for execution. For the duration of the time that the inmate regains his cognitive faculties – there is no need to eradicate them permanently – he is sane enough for execution. Any potential for relapse due to non-treatment is beside the point. Nor does the 'artificiality' of the sanity make any difference. If a psychotic were to commit a crime while on medication – medication that made him aware and in control of his conduct – the artificiality of his sanity would provide no defence. Just as culpability for the commission of a crime depends on the cognitive and volitional state of the actor at the time, and not the cause of the actor's condition, the competence of the death row inmate should not be affected by the basis for that competence (Latzer, 2003, p.9).

In response to what Latzer says about reasons to doubt the effectiveness of such techniques above, I think that the question of whether medication is effective to render condemned inmates *sufficiently* competent for execution in the *majority* of cases is an empirically and conceptually open question. On the *empirical* side, it is at least conceivable that tests could be administered to treated inmates to ascertain whether *most* of them are in fact rendered sufficiently competent by such treatments. And on the *conceptual* side, it is debatable precisely which mental capacities a person must possess, and in what degrees they must possess them, to be *sufficiently* competent. For instance, precisely what mental capacities and in what degrees must a person have to be competent to stand trial, as opposed to being competent for incarceration, or execution, or even something else? These are difficult conceptual and moral questions which Latzer does not address, and which are still a subject of fierce disagreement (e.g. see the collection of papers in Schopp, Wiener, Bornstein, & Willborn, 2009). However, precisely because so much hinges on empirical facts and on contentious conceptual and moral issues, for the sake of argument I will put these matters aside and grant Latzer's assertion.

Furthermore, with Latzer (2003), and against what Horstman (2002) says in the above-quoted passage, I agree that the *impermanence* of artificially-induced sanity should not matter to the question of whether someone is competent for execution. If we presuppose as Latzer and many others do that capital punishment is at least sometimes permissible, then what ought to matter is the persistence of sanity until the condemned inmate's life is extinguished. The fact that their sanity *would have* evaporated once the effects of the drugs wore off if they had not been executed is immaterial, or at least it is far from clear why this should matter.

However, I reject Latzer's argument in the second paragraph quoted above for three reasons. Firstly, contrary to what Latzer asserts, there are compelling reasons to doubt that a person who commits a crime

while on mental capacity restoring medications necessarily ought to be viewed as responsible for that crime (Carter, Ambermoon, & Hall, 2010). One problem here is 'merely' epistemic[13] – i.e. we simply have no way of knowing whether the person who committed the crime while on the medications did so out of their own un-tainted volition, or because the meds did not work, or due to the medications' side effects. In fact, the person themselves may not know whether they committed the crime out of their own un-tainted volition or because of the drugs' (side) effects simply because the true causes of our own actions can be just as inaccessible to us as they are to others (Davies, forthcoming). However, our worries may run much deeper – i.e. they may not be 'merely' epistemic in the above sense. For instance, perhaps there simply is no fact of the matter – i.e. nothing even for the person concerned to know – about whether the crime should be attributed to *them* or to *the drugs*, since what we have post-treatment is not a *person* and their *meds*, but rather we just have a *treated person* and this person is not the same as the person with whom we were faced prior to treatment. Carter and colleagues (2010) provide some very evocative examples which support this claim:

> There are cases, however, in which DRT [dopamine replacement therapy] induces behaviour that individuals claim are authentic. For example, one male who became fascinated with anal sex following DRT claimed that he had these desires prior to DRT treatment but was too embarrassed to act on them. The medication allowed him to 'realise these desires'. His interest in these sexual behaviours stopped following a change in his medication, and he later expressed regret at his behaviour. Similar experiences have been expressed by users of other drugs that affect the dopaminergic system. Singh interviewed adolescents treated for ADHD with Ritalin (a drug which also increases dopaminergic stimulation) and their parents, and found that the attribution of authenticity to various actions depended on whether the behaviour was seen as positive or negative. For example, a child's bad behaviour was often attributed to a failure to take their medication, whereas success on the sporting field was more likely to be attributed to the child. This reflects a common human characteristic documented by social psychologists to take personal credit for our successes and blame circumstances for our failures. There has been no attempt as yet to examine whether DRT patients believe that these changes in behaviour are authentic expressions of who they are or simple neurochemical reflexes (Carter et al., 2010, pp. 5-6).

In the very least, these examples demonstrate the more acute form of the epistemic problem that I discussed above – i.e. that even subjects themselves may vacillate about whether to attribute behaviour to themselves or to the medications that they are on, and not merely because they try to cover up their own

[13] I say 'merely' in order to emphasize the point that although epistemic questions do not undermine the existence of facts about responsibility but only our knowledge of those facts – I am of course assuming here that a realist stance on responsibility is defensible, though I will not argue for this point – never the less we should not underestimate the serious problems that these epistemic problems pose for attempts to settle questions about responsibility.

guilt but rather because they themselves do not know this. However, it is also plausible to interpret the phenomenology and self-reports present in these examples as evidence that what we have prior and subsequent to treatment are two different selves, and that what one of these selves does may not necessarily be attributable to the other self. This is relevant to Latzer's (2003) argument for it shows that it is far from clear, in both a deep metaphysical sense as well as a 'mere' epistemic sense, that a person who performs a crime while on such medications would necessarily be responsible for that crime. And to the extent that this undermines Latzer's claim that *"[i]f a psychotic were to commit a crime while on medication…the artificiality of his sanity would provide no defense"* (2003, p. 9), this also undermines his consequent claim that *"[j]ust as culpability for the commission of a crime depends on the cognitive and volitional state of the actor at the time, and not the cause of the actor's condition, the competence of the death row inmate should not be affected by the basis for that competence"* (2003, p. 9).

Secondly, it could be argued that once one person meddles with another person's mind, the former (the 'manipulator') for ever remains at least partly entangled in what the latter one does and is thus at least partly responsible for it because their agency has become inextricably linked to their patient's agency. This is certainly why responsibility is allegedly undermined in so-called 'manipulation cases' discussed in the compatibilist literature (e.g. Fischer & Ravizza, 1998), and the leading intuitions there are the same as the ones that I draw upon above – namely, that the manipulator takes over at least some of the responsibility from the subject of their manipulation, and that the subject loses ownership of their mental capacities (i.e. they cease to be genuinely or authentically *their* mental capacities). Latzer (2003) is not entitled to suppose that it is irrelevant how (i.e. by what causes) a person came to possess their mental capacities, and so this is another reason to suppose that the fact that a person's sanity is *artificial* is significant after all.

Thirdly, as I have already suggested above, it is far from clear that the mental capacities which a person must have to be competent to stand trial, are the same as the ones required in other contexts – e.g. to be competent for incarceration, or execution, or even to be a fully responsible person in the sense of being a legitimate target of attributions of praise and blame for what they do and of admiration and condemnation for who they are. The literature on competence in legal settings is filled with disagreement about which mental capacities are needed for responsibility and competence at different legal stages – i.e. when the crime is originally committed, when they subsequently stand trial for it, when they are eventually punished, and maybe even when they are assessed for early release by a parole board (e.g. Buchanan, 2006; Burrows & Herbert, 2005; Mossman et al., 2007; Rogers, Blackwood, Farnham, Pickup, & Watts, 2008; Otto, 2009; Winick, 2009). Nevertheless, the single message which comes through loud and clear is that the mental capacities which a person must possess for full responsibility or for competence at these difference stages vary greatly. Consequently, Latzer (2003) is simply not entitled to avail himself of the analogy between the person who commits a crime while on medications, and another who is treated to make them competent for execution, because qualitatively and quantitatively different mental capacities may be required in these two different contexts. Thus, even if it were possible to restore the mental capacities that are required for attributions of responsibility for what one does through direct brain interventions (and I have argued that this

is questionable), it would not necessarily follow that it must therefore also be possible to restore the mental capacities required to make people into legitimate candidates for retributive punishment.

In summary, I discussed two objections to the claim that direct brain interventions could be used to help achieve justice in the traditional sense. Firstly, I looked at the 'mask of sanity' objection which claims that such treatments only make it seem as if the treated person is once again a legitimate candidate to be put on trial or punished. My main point here was that Latzer's (2003) response to this objection assumes way too much – in particular, that the question of whether medications actually are effective to render most condemned inmates sufficiently competent is an empirically and conceptually open one – although I did not press this point any further. Secondly, I also looked at Latzer's (2003) response to the 'artificial sanity' objection, which claims that sanity installed through such treatments cannot help us achieve traditional justice. Here I offered two reasons to reject Latzer's claim that artificial sanity can substitute for normal sanity – one drew on examples cited by Carter and colleagues (2010), and the other drew on the compatibilist literature about responsibility in manipulation cases – and I also argued that even if it were possible to restore the mental capacities needed for full responsibility, that would not entail that it is possible to restore the mental capacities required for competence to be punished or to stand trial.

Mental capacity restoration and therapy

There are at least three problems with the claim that direct brain interventions can help further the aim of therapeutic justice.

Firstly, as I have already argued above, once we meddle with a person's mind to treat their irresponsibility or non-responsibility – i.e. to make them responsible (as opposed to non-responsible) – paradoxically they may never again be fully responsible for anything that they do. Although such treatments aim to restore people's status as moral agents – as individuals whom we can praise and blame for what they do – if we accept the leading intuitions in the previously cited compatibilist literature on responsibility in manipulation cases, then we may again be forced to conclude that once one person meddles with another person's mind, the latter person will never again be fully responsible for what they do since at least some of their responsibility will be transferred back to the former person (i.e. to the 'manipulator').

Secondly, in the fiction novel 'A Clockwork Orange', Anthony Burgess (2000) writes the following about Alex, the hyper-violent young protagonist who is sentenced to undergo a treatment intended to make him into a responsible individual in the sense of being a legitimate target for admiration and condemnation for who he is: *"Does God want goodness or the choice of goodness? Is a man who chooses the bad perhaps in some way better than a man who has the good imposed upon him?"* (Burgess, 2000, p. 71).

Burgess' point, if I get him right, is that reform which is forced or inflicted upon one person by another has little if any value — i.e. that moral virtue of this sort is not something that can be implanted into people by brute force. If Burgess is right then the very aim of making people responsible in this sense of the word would be conceptually flawed at a very fundamental level, because unless one already is responsible or chooses of one's own untainted volition to become responsible, then no amount of other-inflicted changes could ever *make* one into a responsible person. At best, one might become a well-behaved puppet or

automaton, or perhaps a prisoner of the 'do gooder' impulses that have been implanted in one's psyche like ever-watchful and ever-knowing security guards, but puppets and automatons are simply not subjects of moral admiration or contempt.

And thirdly, there is also the worry that such treatments would not restore mental capacities but rather alter character or personality. Consider for instance a person who is callous, uncaring and feels little if any empathy for others, as a consequence of which they are always flying off the handle and beating people up. Is it more accurate to view such a person as someone who has a capacity deficit (e.g. they have diminished capacity for self-control and diminished affective capacities), or as someone who has a set of nasty and maybe even condemnable character flaws (i.e. they are an unsavoury individual with a mean streak, a nasty temper and a short fuse)? Should such a person be treated for their capacity deficits, or despised for their character flaws? As I have argued elsewhere, the answer is far from clear, and it hinges not merely on empirical facts but at least to some extent also on irreducibly conceptual and moral issues (Vincent, 2011a). Furthermore, as some of the literature on psychopathy has noted, *prima facie* both of these descriptions seem apt — what looks like madness when it is viewed from one angle, can also be viewed as badness when it is from another angle (e.g. Maibom, 2008; Reimer, 2008; Sadler, 2008). But if the same state of affairs equally admits of two radically different interpretations – viewed one way it is a rabbit, but viewed another it is a duck – then how are we to decide whether to treat such people's illness/madness or to despise them for being evil/bad?

There are two reasons why this question matters. Firstly, as Lewis (1963) has argued, without a way of distinguishing ducks from rabbits – or madness from badness in the case at hand – in the name of therapy and compassion such treatments might instead allow the state to inflict unspeakable brutality onto citizens whose character or values it disapproves of:

> [I]f crime and disease are to be regarded as the same thing, it follows that any state of mind which our masters choose to call 'disease' can be treated as crime; and compulsorily cured. [But] one school of psychology already regards religion as a neurosis[, and w]hen this particular neurosis becomes inconvenient to government, what is to hinder government from proceeding to 'cure' it? [W]hen the command is given, every prominent Christian in the land may vanish overnight into Institutions for the Treatment of the Ideologically Unsound. [T]he Humanitarian theory of punishment ... carries on its front a semblance of mercy which is wholly false (Lewis, 1963, p. 229).

Secondly, it also matters because if the 'bad character' interpretation is more appropriate in some cases than the 'capacity deficit' interpretation, then in those cases being sentenced to treatment would be a macabre death sentence in disguise — a death of personality (Greely, 2008). In some cases (those which involve mental capacity deficits) therapy might indeed be achieved, but in others (i.e. those which involve character flaws) brutality not therapy is all that would be achieved. And if it should turn out that there is not even a valid distinction here to be drawn between mental incapacities and character flaws – for instance, if it should turn out that the neurological correlates of our mental capacities are identical to the neurological correlates of our character – then all such treatments would effectively also modify people's character or personality.

Admittedly, nothing that I have said above has shown that what such treatments would do is alter character or personality rather than treat mental incapacities, and thus I cannot claim to have positively established that such treatments would fail to achieve their therapeutic aims. However, I need not show this for my argument to meet its target, because all I need to show is that we have ample grounds to worry that we currently have no way to distinguish character flaws from capacity deficits, and thus that to be on the safe side we should abstain from 'treating' people with direct brain interventions until we have gathered more empirical data on this topic and analysed the conceptual basis of the distinction between capacity and character.

In summary, I have argued that there are three problems with trying to further the aim of therapeutic justice by using direct brain intervention based techniques for mental capacity restoration. Firstly, by meddling with a person's mind we might (rather paradoxically) rob them forever of the opportunity to be fully responsible for what they do. Secondly, although direct brain interventions might succeed in making one into a well-behaved puppet or automaton, they cannot make one into a genuinely responsible person since puppets and automatons are not the subjects of moral admiration or contempt. Thirdly, far from being merciful and compassionate, such treatments may inflict the worst forms of brutality upon their victims by inflicting personality and character changes upon them — i.e. it might turn out that what such treatments do is to alter character rather than mental capacities.

A problem with capacitarianism?

The main point advanced by the previous section's arguments can be stated in two different though equivalent ways. Stated one way, these arguments show that direct brain interventions cannot help us to achieve justice in either the traditional or the therapeutic sense. Stated another way, they show that direct brain interventions cannot help us to assess the responsibility of someone who becomes mentally ill subsequent to committing their crime or to hold them responsible, to expect them to take responsibility for what they did, or to make them fully responsible and maybe even less irresponsible.

The first way of stating the point advanced by the previous section's arguments is already interesting enough, for it challenges practices (and their rationale) which are already affecting people's lives today. For instance, it challenges the practices of medicating defendants and condemned inmates so that they can be tried and/or punished, and the administration of cyproterone acetate and SSRIs to convicted violent offenders to treat the causes of their criminality.

However, what makes the second way of stating the previous section's point particularly interesting is that it raises the possibility that capacitarianism might be false, flawed or at least limited. By showing that responsibility does not track *restored* mental capacities, the previous section in the very least seems to rein in the scope of the central capacitarian thesis. But given the pervasive role that capacitarianism plays in our thinking about responsibility, this should be a ground for concern, for if responsibility does not track mental capacity across the board, then how can we be sure that it even tracks mental capacity in the more 'garden variety' cases?

A brief defence of capacitarianism

When a foundational assumption of this sort which informs how we reason about moral issues is challenged, one way to proceed is to conclude that this assumption must indeed be false — i.e. that the central hypothesis expressed by this assumption has been falsified. In the case of capacitarianism, this would involve concluding that the central capacitarian thesis that responsibility tracks mental capacity must after all be incorrect.

However, a less drastic option might be available – namely, to check whether something else (an auxiliary assumption or set of assumptions) can explain why the central assumption seems to generate or support conclusions that clash with reflective moral intuitions – and that is indeed precisely what I will attempt to do below. Firstly, I will argue that capacities of all sorts, not just mental ones, affect our responsibility judgments. Secondly, I will explain why capacities seem to matter to responsibility. Thirdly, I will list some other things that also affect our responsibility judgments. And finally, I will end by explaining how all of this bears on the above threat to capacitarianism and on whether justice (in either the traditional or therapeutic sense) might be furthered along at least in some cases by using direct brain interventions.

Not just mental capacities

Imagine that a child drowns at the beach, and we ask whether a person who was on the shore looking out across the waves in the child's direction (the onlooker) is responsible for their drowning because they did nothing to save them. There are several ways in which a mental incapacity could absolve them of responsibility, even if only partially, for the child's drowning. For instance, suppose that we learn that the onlooker was actually looking out across the waves at the birds that had gathered above the water, but that they had no inkling of what was attracting the birds' attention — i.e. a child struggling for life in the water. Alternatively, suppose that the onlooker had noticed the drowning child, but that they have a pathological and paralysing fear of water — a fear which they were in fact trying to overcome by taking a stroll to the beach, only to be traumatised even further by this awful turn of events. Or even suppose that the onlooker suffered from a mental illness due to which they could not think clearly.

In each of these cases a *mental* incapacity is what would undermine the onlooker's responsibility – respectively, a lack of knowledge, a volitional impairment, and a cognitive impairment.[14]

However, physical incapacities could also absolve the onlooker of responsibility. For instance, the onlooker might not know how to swim, in which case we may not be justified in claiming that they had a

[14] A lack of knowledge is usually classified as a cognitive impairment, though see following note and the related text about (e.g.) knowing how to swim, which I classify as a physical incapacity rather than a mental incapacity.

responsibility to save the child in the first place which they subsequently breached by doing nothing to save them.[15] Alternatively, maybe they knew how to swim but they lacked the physical strength or stamina required to swim out towards the child – perhaps they had only just swam back ashore themselves and were too worn out and exhausted, or maybe they were sitting there in a wheel chair paralysed from the neck down. The physical tools at our disposal can also affect our capacities – for instance, if the scenario had been such that in the vicinity was a life boat, a rope or a lifebuoy/ring which could have been thrown to the child, then that again might have extended the onlooker's capacities which in turn might have justified the supposition that they had a responsibility to save them. These are just a few examples of how *non-mental* capacities might be relevant to responsibility.

My point here is simply that it is plausible that responsibility may co-vary with or track a wide range of different capacities, many of which are not necessarily mental.

Why capacities matter

Capacitarianism's normative appeal seems to derive from the maxim that *ought* implies *can*. The links here are between our *responsibilities* and what we *ought* to do on the one hand, and our *capacities* or what we *can* do on the other hand. However, there are at least two ways of explaining how our capacities (what we can do) might have a bearing on our responsibilities (what we ought to do).

In the positive version of this explanation, capacities *generate* responsibilities. The idea here is that we ought to do what we have most reason to do, and what we can and cannot do (along with a range of many other things, as I explain below) generates the reasons that we have to do various things. An inference is thus first made from what capacities I have to what I have reason to do, and then another inference is made from what I have most reason to do to what I ought to do – i.e. we move from *capacity* claims via *reasons* claims to *ought* claims. On this account, if I cannot save a child from drowning – perhaps because I do not know that they are drowning, or because I cannot swim, or because I do not have a rope to throw to them – then it is simply not true that I ought to save them (unless I am responsible for the fact that I cannot do this – see the discussion of the role of history below). The reason why I would not be blameworthy for not saving them is because I was not in the first place even subject to that saving duty. On the other hand, in the negative explanation capacity *regulates* duties. The idea here is that regardless of the source of our duties,

[15] Two points should be noted. Firstly, I am treating the onlooker's lack of knowledge of how to swim as a physical incapacity because it is an instance of 'knowing how' rather than 'knowing *that*', and because it involves a physical activity. In a sense, imagine that the onlooker says "But I can't swim." To me this seems natural to view this as a physical incapacity. Secondly, the onlooker may not be absolved of responsibility if they should have had the capacity to swim — I return to this point below.

on this second view our incapacities can excuse departures from those duties. On this latter account, the three cited considerations – i.e. I do not know that the child is drowning, I cannot swim, or I have no rope – do not extinguish the saving duty, but rather they provide an excuse for departing from it. The reason why I would not be blameworthy on this second account is because although I did have the saving duty, my incapacity provided an excuse for departing from it.

Two advantages of the negative explanation are that only it has an explicit place for excuses and justifications which play a prominent role in much ordinary and legal thinking about responsibility, and arguably it also more adequately captures the rich structure of practical reasoning in which some considerations discount, undermine and invalidate (rather than just outweigh or extinguish) other considerations. Nevertheless, I suspect that both views of the relationship between capacity and responsibility will generate the same responsibility judgments, and since I find the positive explanation simpler, in what follows my discussion will be framed in terms of it rather than in terms of the negative explanation. Thus, the idea is not to read off a person's responsibilities simply from an assessment of their capacities (this would assume that *can* implies *ought*), but it is rather that in determining what responsibilities a person has we should, among other things, consider what capacities they (ought to)[16] possesses – i.e. the idea is that *can*, taken together with a range of other considerations, implies *ought*.

What else matters

Although capacities play an important role in informing responsibility claims, they are not the only things that matter to responsibility. For instance, to determine what a person is responsible for and the degree of their responsibility for it, it surely also matters *what that person did* (e.g. killed another person, stole an item, offended someone, mildly bruised another's ego, etc) and the degree to which their actions causally contributed to the outcome (e.g. did they play a crucial *sine qua non* role in bringing it about, were their actions just one of a number of necessary contributions, or were their actions a dispensable 'overkill' contribution without which the outcomes would still have come about, etc).

Historical factors also play an important role. For instance, in the above example we would assess the onlooker's responsibility differently if they were blameworthy for their own incapacity – perhaps because they should have learned how to swim, or because they should not have allowed their energy levels to drop like that. As Smith points out, when an agent's incapacity is caused by *"an initial [benighting] act, in which the agent fails to improve (or positively impairs) his [own] position"* (Smith, 1983, p. 547), the exculpatory value of that incapacity is itself diminished and maybe even extinguished.

[16] I discuss the need for this qualification below.

A vast range of *normative and policy considerations* can also affect our responsibility judgments. For instance, suppose that the water at the beach was teaming with hungry sharks; under such circumstances it would surely be unreasonable to insist that the onlooker should have swam out in an attempt to save the child since this would be tantamount to expecting them to almost certainly commit suicide — that would be too much to expect of anyone. On the other hand, if the onlooker had previously undertaken to care for the child at any cost, it might be reasonable to blame them for not trying to save the child then if we deem this to have been a binding undertaking on their part. Similarly, we might have sound policy reasons to treat children's guardians as responsible for the children's actions and for what happens to those children. By clearly delineating people's responsibilities, and in this case by imposing them onto those who are in the best position to take care of the child, we are more likely to avoid a situation where everyone assumes that it is somebody else's responsibility to look out for the child's welfare.

Normative considerations might also play a role in determining *how much* of a given capacity a person needs in order to have a responsibility to do something – i.e. normative considerations might play a role in setting the threshold of capacity required for responsible agency in a given context or sphere – and precisely how a person should be treated in order to be appropriately held responsible for what they have done – e.g. whether the right punishment for theft is 5, 10, 15 or 20 lashes of the whip, amputation of the hand that stole the item, incarceration (and if so, for how long), execution, or something else entirely.

And finally, *personal identity* considerations also seem pertinent to responsibility assessments, since at least prima facie a person who possesses all of the right mental capacities might still act while not being themselves, and then it would seem inappropriate to attribute the actions to them. As I already mentioned, this is the very point of Fischer and Ravizza's (1998) ownership condition, which specifies that to be responsible for an action (or its outcome), that action must have issued from *our own* moderately reasons-responsive mechanism.

My point is that responsibility claims are affected by a wide range of considerations and not just by what capacities the person to whom the claim pertains possesses. Above I mentioned the person's actual behaviour, their causal contribution to the outcome, historical factors, their obligations, normative and policy considerations as well as personal identity considerations. Hence, the central capacitarian thesis that responsibility tracks mental capacity should be understood against the general background of these other auxiliary assumptions.

Concluding remarks

The above discussion helps explain why it is not yet time to start writing an eulogy for capacitarianism.

Given the role that historical considerations play in informing responsibility judgments, it should come as no surprise that *artificial sanity* will not suffice for seeing to it that justice is done – i.e. to re-enable mentally ill defendants and condemned inmates to stand trial and to be punished, so that their responsibility can be assessed, so that they can be held responsible, and so that they can take responsibility for what they have done. After all, the particular historical trajectory that leads these parties to possess those particular mental capacities undermines the claim that those are genuinely *their* mental capacities. Put a different way, certain kinds of histories undermine the ownership condition, and that in turn undermines responsibility. Thus, the reason why responsibility would not track mental capacity in cases of forced mental capacity restoration is because the historical and ownership conditions discussed above would not be met – i.e. this is not a counter-example to capacitarianism because historical and ownership considerations are among the auxiliary assumptions which provide legitimate exceptions to the general capacitarian rule. The *artificiality* of the sanity is thus significant not qua it having been brought about through *unnatural* methods, but rather qua being *inflicted* and qua being inflicted *in a manner* which gives us no choice over whether and how we will be changed. This is from whence at least some of our reservations about medicating people so that they can stand trial and be punished come.

The role that normative considerations play in informing responsibility judgments also helps to explain why we might feel reserved about whether a particular form of competence or responsibility is restored by a particular treatment. For instance, depending on what we think is the purpose of punishment – for instance, retribution, deterrence, reform, or expression of solidarity with victims and their families – defendants may need a different set of mental capacities (and in different degrees) to be competent for punishment because the mental capacities one might need to be reformed may differ radically to the mental capacities one might need to have in order for victims and their families to get a sense of closure from seeing one punished. And for identical reasons, just because we might be able to restore the mental capacities required for one context (e.g. standing trial where responsibility is assessed and taken (or not) by the defendants), that does not mean that we must also be able to restore the mental capacities required for another context (e.g. punishment where people are held responsible and take responsibility).

What I said above about the role of history (that certain ways of acquiring mental capacities fail to satisfy the ownership condition) also explains why direct brain intervention techniques might face an up-hill battle in trying to help us achieve the therapeutic aim of making people into fully responsible moral agents –

i.e. into people who could be released back into society and be legitimate targets for attributions of praise and blame for anything that they do thenceforth. Part of the problem is that if people are treated involuntarily – i.e. against their will – then the resulting capacities that they might acquire will not be *their own* and thus their possession of those capacities will not confer 'fully responsible' status onto them. But even if the treatments are consented-to or asked-for,[17] any direct brain interventions which implement large-scale changes in one fell swoop would most likely break sufficient continuity between the pre-treatment and the post-treatment person, which in turn would give us grounds for concern that the post-treatment person is not in an authoritative position to retroactively endorse those changes. After all, the person who would be conducting the re-assessment might be too different from the person who they were prior to the treatment, and so any endorsement that they give might be insufficiently *partial* to warrant the claim that *they* endorse those changes in retrospect.[18]

The above reflections suggest that the celebrated effectiveness of direct brain interventions is also probably their down-side – i.e. because when too much is changed in one single step, this removes the treated person's authority to retroactively endorse what has been done to the pre-treatment person. The difference between direct brain interventions and more conventional techniques for changing people, such as cognitive behavioural therapy or even some of the positive effects that we hope to achieve when we incarcerate someone, would thus appear to be that with the latter methods the treated person always retains the ability to stick to their guns and resist being changed. On the other hand, direct brain interventions by-pass the treated person's ability to veto the changes that are being made to them – i.e. they slide in undetected 'under the radar', robbing them of the opportunity to resist but also to endorse those changes – and the greater the change that is involved, the less likely it is that the treated person will be in a sufficiently authoritative position to retroactively endorse those changes. However, this also leaves open the possibility of using direct brain interventions to make many small and preferably reversible changes, all along allowing the treated individual to say "Stop there, let's go back – I no longer endorse what's being done to me!" Effectiveness in small doses might be acceptable, and so perhaps this might be taken to suggest that the

[17] And even if we put aside worries about whether genuine consent can be given in 'It's either prison or treatment for you!" style conditions. See Bomann-Larsen (2011) for a discussion of some problems with voluntary treatments.

[18] In Chapter 8 of *Responsibility and Control*, Fischer and Ravizza (1998) outline their views about *"the process by which a mechanism leading…to an action, becomes one's own"* (p. 207). They argue that there are certain processes through which a person can *retroactively* come to own the mental capacities which they have come to possess, and they call these processes 'taking responsibility' (I use this expression to mean something different). However, although I find most of their compatibilist theory to be very compelling, for reasons outlined briefly in this paragraph I do not endorse what they say about this matter.

sorts of direct brain intervention techniques that we should be working on – if we should be working on any such techniques, that is – are ones that would only make small and reversible changes.

Finally, nothing that I have said in this last section substantially affects my earlier claims about whether direct brain interventions could be used to make people more responsible in the sense of making them into legitimate candidates for admiration and condemnation for *who they are*. Whatever value there might be in being a responsible person in this sense is not something that can be inflicted upon us from the outside by others. Furthermore, I would still worry about changing people's character through direct brain interventions even if the changes made were small, gradual and reversible, and this highlights an interesting difference between making people more responsible in the 'legitimate candidates for praise and blame for what they do' sense and making them more responsible in the 'legitimate candidates for admiration and condemnation for who they are' sense.

References

Aristotle (1976). Book Three: moral responsibility — two virtues. The Ethics of Aristotle. In J. Barnes & H. Tredennick (Eds.), (111-141). London: Penguin Books.

Atkins v. Virginia (2002).

Bomann-Larsen, L. (2011). Voluntary rehabilitation? On neurotechnological behavioural treatment, valid consent and (in)appropriate offers. *Neuroethics*. DOI: 10.1007/s12152-011-9105-9.

Bradford, J. M. W., & Pawlak, A. (1993). Double-blind placebo crossover study of cyproterone acetate in the treatment of the paraphilias. *Archives of Sexual Behavior, 22(5)*, 383-402.

Buchanan, A. (2006). Competency to stand trial and the seriousness of the charge. *The Journal of the American Academy of Psychiatry and the Law, 34(4)*, 458-465.

Burgess, A. (2000). A clockwork orange (with an introduction by Blake Morrison). London: Penguin Books.

Burrows, M. S., & Herbert, P. B. (2005). Competence to stand trial does not conclusively equate to competence to waive trial counsel. *The Journal of the American Academy of Psychiatry and the Law, 33(4)*, 557-558.

Carter, A., Ambermoon, P., & Hall, W. D. (2010). Drug-induced impulse control disorders: A prospectus for neuroethical analysis. *Neuroethics*. DOI: 10.1007/s12152-010-9071-7.

Chatterjee, A. (2007). The promise and predicament of cosmetic neurology.. In W. Glannon (Ed.), *Defining right and wrong in brain science* (302-311). New York: Dana Press.

Connecticut v. Kenneth Curtis (1999).

Davies, P. S. (forthcoming). Skepticism concerning human agency: sciences of the self vs. 'voluntariness' in the law. In N. Vincent (Ed.), *Legal responsibility and neuroscience*. New York: Oxford University Press.

Dawkins, R. (2006). *Let's all stop beating Basil's car*. Retrieved from http://www.edge.org/q2006/q06_9.html - dawkins.

Dennett, D. C. (1984). *Elbow room: the varieties of free will worth wanting*. Cambridge, MA: MIT Press.

Duff, R. A. (2005). Punishment, dignity and degradation. *Oxford Journal of Legal Studies, 25(1)*, 141-155.

Eisenberg, L. (2004). Medicating death row inmates so they qualify for execution. *Virtual Mentor, 6(9)*.

Farah, M. J. (2002). Emerging ethical issues in neuroscience. *Nature Neuroscience, 5(11)*, 1123-1129.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge, UK, Cambridge University Press.

Ford v. Wainwright 477 U.S. 399 (1986).

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy, 66(3)*, 829-839.

Glannon, W. (2002). *The mental basis of responsibility*. Aldershot: Ashgate Publishing.

Greely, H. T. (2008). Neuroscience and criminal justice: Not responsibility but treatment. *University of Kansas Law Review, 56*, 1103-1138.

Hart, H. L. A. (1968). *IX. Postscript: Responsibility and Retribution. Punishment and Responsibility*. Oxford: Clarendon Press.

Horstman, L. A. (2002). Commuting death sentences of the insane: A solution for a better, more compassionate society. *University of San Francisco Law Review, 36*, 823-852.

Latzer, B. (2003). Between madness and death: the medicate-to-execute controversy. *Criminal Justice Ethics, 22(2)*, 3-14.

Lekovic, G. P. (2008). Neuroscience and the law. *Surgical Neurology, 69*, 99-101.

Lewis, C. S. (1963). The humanitarian theory of punishment. *Res Judicatae, 6*, 224-230.

Macey, J. (2010). Antidepressants may help violent offenders. ABC News Online. Retrieved from http://www.abc.net.au/news/stories/2010/12/07/3087359.htm.

Maibom, H. L. (2008). The mad, the bad, and the psychopath. *Neuroethics, 1(3)*, 167-84.

Mossman, D., Noffsinger, S. G., Ash, P., Frierson, R. L., Gerbasi, J., Hackett, M., Lewis, C. F., Pinals, D. A., Scott, C. L., Sieg, K. G., Wall, B. W., & Zonana, H. V. (2007). AAPL practice guideline for the forensic psychiatric evaluation of competence to stand trial. *The Journal of the American Academy of Psychiatry and the Law, 35(4)*, S3-S72.

Otto, R. K. (2009). Meaningful consideration of competence to be executed. In R. F. Schopp, R. L. Wiener, B. H. Bornstein & S. Willborn (Eds.), *Mental disorder and criminal law* (191-204). New York: Springer.

Reimer, M. (2008). Psychopathy without (the language of) disorder. *Neuroethics, 1(3)*, 185-198.

Riggins v. Nevada 504 U.S. 127 (1992).

Rogers, T. P., Blackwood, N. J., Farnham, F., Pickup, G. J., & Watts, M. J. (2008). Fitness to plead and competence to stand trial: a systematic review of the constructs and their application. *The Journal of Forensic Psychiatry & Psychology, 19(4)*, 576-596.

Sadler, J. Z. (2008). Vice and the diagnostic classification of mental disorders: A philosophical case conference. *Philosophy, Psychiatry & Psychology, 15(1)*, 1-17.

Sapolsky, R. M. (2004). The frontal cortex and the criminal justice system. *Philosophical Transactions of the Royal Society of London, 359*, 1787-1796.

Schopp, R. F., Wiener, R. L., Bornstein, B. H., & Willborn, S. L. (2009). *Mental disorder and criminal law*. Dordrecht: Springer.

Sell v. U.S. 539 U.S. 166 (2003).

Sher, G. (2009). *The searchlight view. Who knew? Responsibility without awareness*. New York: Oxford University Press.

Singleton v. Norris 319 F.3d 1018 (8th Cir. 2003).

Tancredi, L. R. (2005). *The bad and the mad. Hardwired behavior: What neuroscience reveals about morality*. Cambridge: Cambridge University Press.

Tovino, S. A. (2007). Functional neuroimaging and the law: Trends and directions for future scholarship. *The American Journal of Bioethics, 7(9)*, 44-56.

Vincent, N. (2010). On the relevance of neuroscience to criminal responsibility. *Criminal Law and Philosophy, 4(1)*, 77-98.

Vincent, N. (2011a). Madness, badness and neuroimaging-based responsibility assessments. In M. Freeman (Ed.), *Law and neuroscience, current legal issues* (79-95). Oxford; Oxford University Press.

Vincent, N. (2011b). A structured taxonomy of responsibility concepts. In N. Vincent, I. van de Poel & J. van den Hoven (Eds.), *Moral responsibility: beyond free will and determinism*. Dordrecht: Springer.

Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Cambridge, MA: Harvard University Press.

Winick, B. J. (2009). Determining when severe mental illness should disqualify a defendant from capital punishment. In R. F. Schopp, R. L. Wiener, B. H. Bornstein & S. Willborn (Eds.), *Mental Disorder and criminal law* (45-78). New York: Springer.

Section B: Legal issues of using neuroscience in court

Chapter 3
Legal admissibility of suitable fMRI based lie detection evidence in German criminal courts

Stefan Seiterle
European University Viadrina Frankfurt (Oder)
Faculty of Criminal Law
✉ sseiterle@europa-uni.de

Abstract Discussion about 'lie detection' has almost exclusively been linked with the polygraph. Its use has always been controversial. Yet undoubtedly there is a need for search of truth in trials, considering the unreliability of eyewitness testimony and the subjective standpoints of both judges and juries. In recent years, there have been immense efforts to develop new ways to detect deception, and one of them is using functional Magnetic Resonance Imaging (fMRI). It is too early to tell if this method is more suitable than the polygraph (the denying of admittance of fMRI lie detection for reasons of unreliability by a Magistrate Judge in the Western District of Tennessee in May 2010 had a lot to do with the specific facts of that case and cannot be generalised). Its potential problems must be researched as soon as possible. In 1998, the *Bundesgerichtshof* (German Federal Court of Justice) held that the polygraph must not be used in German courtrooms, but for the first time, the Court explained the prohibition with something other than purely legal arguments. Instead, the polygraph was barred because of alleged *"thorough unsuitability"*. This leads to the question whether a more suitable method – such as, for example fMRI – would now find access to German criminal courts, or whether the (purely) legal arguments that formerly applied against the polygraph could gain relevance again. Even with the consent of the accused, one might still argue that the use of a lie detection method in criminal courts violates the person's human dignity or his or her personal rights. Under which conditions would the consent be valid, if at all? And would admitting deception detection in the courtrooms not violate the accused's right against self-incrimination? Does it make a difference that fMRI looks right into the brain of the examinee? Does this mean it looks into the 'soul' of the accused, as was claimed by the *Bundesgerichtshof* in its 1954 polygraph judgement?

Keywords neuroscience, fMRI, lie detection, admissibility, German criminal law

Introduction

In May 2010, a Magistrate Judge in the Western District of Tennessee ruled that at least the fMRI (functional Magnetic Resonance Imaging)-based lie detection test at issue was inadmissible at trial as it was judged as an unreliable method of determining the defendant's truthfulness (US vs Semrau, 2010).[19] In the

[19] Especially in the German discussion the terms 'lie detection' and 'detection deception' are much critized for it is

same month, the Indian Supreme Court was of the same opinion regarding a 'brain mapping' test with which it was allegedly possible to find out whether the subject has *"experiential knowledge"* (Smt. Selvi & Ors. vs. State of Karnataka, 2010). Despite debate over the polygraph test's suitability, it is used by law enforcement in many countries in the world, either during preliminary proceedings or at trial or both (Vrij, 2008; Honts, 2004). However, in most countries the use is prohibited. In Europe, only very few countries – like Poland[20] – allow polygraph lie detection in criminal procedure. Courts might still perceive neuroscience-based lie detection as unsuited for the purpose of assessing the reliability of the accused's statement. However, after more than a decade of intense research in this field[21] and numerous attempts to introduce neural lie detection in court settings, there is no doubt that we are not dealing with science fiction any more. The day might come sooner than expected that judges will be satisfied with the scale of reliability and validity that a neuroscience-based test provides. Always bearing in mind that humans are more than poor at detecting lies, that witness testimony is among the most unreliable kinds of evidence and that there is yet a strong need for the ascertainment of truth in criminal trials, it is important to examine the (purely) *legal* admissibility of a suitable neuroscience-based lie detection test.

This article will confine itself mainly to the situation in Germany – although a few comments on the discussion in the USA will be made – but many considerations are of general validity, as those about the voluntariness of the accused's consent, the legal 'nature' of brain activity and the self-incrimination clause. The other restriction concerns the area of use: the article will be about the admissibility of neural lie detection in *criminal courts* only. It will also almost exclusively deal with the issue of *legal* admissibility. The crucial and at least in some legal systems also fiddly question *how* the suitability of neuroscience-based lie detection could be ascertained will only be examined in passing, as the sole criterion for admissibility under this aspect in German law is that the evidence at issue is not *"thoroughly unsuited"* (§ 244 s. 3 ss. 2 var. 4 German Criminal Procedure Code (*Strafprozessordnung, StPO*)) – which is a more generous standard than the *Frye* or the *Daubert* standard in US law.

Lie detection applying the polygraph – questioning techniques

A lot has been written about lie detection with the polygraph. Hence, I will not repeat its history again[22]

---

undisputable that none of the existing 'lie detection' techniques are actually measuring lies directly. What it does instead is attempt to interpret data that are gained by substracting the measured activation in one state from a baseline state. This baseline needs to be established for each tested individual (Langleben, Dattilio, & Guthei, 2006). In this article, for simplification, the term 'lie detection' will be used nonetheless.

[20] Art. 192a § 1 ss. 2, 199a Polish Criminal Procedure Code.

[21] For an overview, see Spence (2008), Vrij (2008) and Bhatt, Mbwana, Adeyemo, Sawyer, Hailu, & VanMeter (2009).

[22] For an overview of the history oft he polygraph, see e.g. Ford (2006) and Steller (1987).

and the technique will be outlined only briefly. The polygraph measures physiological functions such as heart rate, breathing rate and skin conductance. But it is not the machine that makes the differences, it is the questioning techniques. There are two main approaches: One does indeed test whether the tested person answers truthfully or not when asked a specific question related to a 'crime'. The most popular questioning technique in this field is the Control Question Test (or Comparison Question Test, CQT). This test is widely used in the USA and other countries. The CQT compares physiological responses to relevant questions (e. g. 'Did you steal the bike?') with responses to irrelevant questions (e.g. 'Is your name Cornelius?') and with responses to particular comparison questions (e.g. 'During the first 18 years of your life, did you ever steal anything?') (Ganis, Kosslyn, Stose, Thompson, & Yurgelun-Todd, 2003). The other method searches for *knowledge*. This paradigm is based on the assumption that an increased psychophysiological response to a question about a specific detail of a particular crime (e. g. the colour of a stolen bike) indicates 'guilty knowledge' and thus involvement in the crime in question (Guilty Knowledge Test, GKT) (Ganis et al., 2003). It is obvious that the logic of the latter test is based more on cognitive reactions of the examinee, whereas the former seems to measure in effect emotional states like guilt or fear. However, it is important to mark that there is still no justifying basic theory that could provide the *reason* for the enhanced reaction with the CQT. In other words: Even the supporters of the CQT are unable to explain what actually causes the stronger reaction to the relevant questions (Rill, 2001); it could be fear or stress, but also cognitive reactions like comprehension, memory recall, response inhibition, etc. According to Ganis et. al. (2003), there is at least one general problem with the CQT: This polygraph method detects increases in measures that reflect increased arousal, which is at least typically, as mentioned, interpreted as reflecting guilt and fear. These measures can confuse lie detection in two ways. First, guilt and fear can occur in many situations other than during deception, which is why the results do not necessarily indicate deception as such. Second, if the deceptive person does not feel guilty or is generally 'cold-blooded', he or she may not show the physiological reaction (Ganis et al., 2003).

New approach: fMRI-based lie detection

The three neuroimaging technologies/modalities mostly used to measure brain activity are functional Magnetic Resonance Imaging (fMRI), electroencephalography (EEG), and positron emission tomography (PET). fMRI uses a MRI scanner to measure active brain blood flow, EEG uses electrodes attached to the scalp to measure electrical activity, and PET measures the absorption of small amounts of radioactive materials introduced into the subject's body (Moreno, 2009). As for the question of legal admissibility, there is no difference between whether fMRI or EEG is used. However, there is an important difference between those two and PET: PET requires an intravenous line to be placed into the subject. This method is, unlike the

others, *invasive* (Moreno, 2009). At least under German law, that makes it a prohibited interrogation method *per se* (given that brain activity is regarded as testimonial, which is the case, see below, § 136a s. 1 of the German Criminal Procedure Code (StPO)). Although there has been research on lie detection with PET[23], its likelihood to find access to the courts is definitely smaller than is the case with EEG and fMRI.

There are more than 20 peer-reviewed scientific articles that deal with experiments on lie detection with fMRI. Research on lie detection with EEG seems to be far poorer. This article will therefore not be about the (in)famous Farwell's 'brain fingerprinting'[24] and other attempts with EEG – although it cannot be said that this technology does not have potential for lie detection purposes – but will concentrate on fMRI. As with polygraph lie detection, researches have not found a 'specific lie response' when using fMRI for lie detection, and it is doubted that there is a brain pattern that is a singular sign for deception (Vrij, 2008). However, the studies indicate that the brains of deceptive subjects are active in other areas than when the subjects are being truthful. Although these brain areas are not the same in each study, there is some evidence that the lateral and medial prefrontal cortex play an important role in deception (Abe et al., 2006; Ganis et al., 2003; Langleben et al., 2005; Priori et al., 2008). Such activation has been associated with memory-related and executive control processes (Christ, Van Essen, Watson, Brubaker, & McDermott, 2009). According to Ganis et al. (2003), it seems that different types of lies are at least partly modulated by different neural substrates (see also Abe et al., 2006; Priori et al., 2008). There are just a few fMRI studies yet that aimed at detecting deception or concealed knowledge at the *individual* level (Davatzikos et al., 2005; Ganis, Rosenfeld, Meixner, Kievit & Schendan, 2011; Kozel, Johnson, Mu, Grenesko, Laken, & George, 2005; Kozel et al., 2009; Langleben et al., 2005; Nose, Murai, & Taira, 2009). They have revealed accuracy rates of up to 90%.

However, this article is not the place to go into a detailed description of the relevant studies or the methods and techniques applied, as its purpose is to look into the *legal* admissibility of a method that is at least not unsuited for lie detection (see below). Further discussion about more scientific issues can be found in the relevant literature (see above; see also e.g. Greely & Illes, 2007).

Suitability of scientific evidence

In particular in the USA, but also, for example, in India, there has been intense discussion about which requirements have to be met to determine a method as scientifically valid. As mentioned above, the studies about fMRI lie detection have shown up to 90% correct classifications of truthful/deceptive answers. However, many authors mention issues that raise concern in terms of a lack of reliability (the consistency of the measurement), construct validity (do studies test what they purport to test?) and external validity (do

[23] Mainly in Japan, see Abe et al. (2009), Abe, Suzuki, Mori, Itoh, & Fujii (2007) and Abe et al. (2006).

[24] See e.g. Stoller & Wolpe (2007).

laboratory results predict real-world outcomes?) (Schauer, 2010). These issues include the small number of studies with individual effects, the lack of replication, the small and non-diverse groups of subjects, the inconsistency of reported regions of activity, the artificiality of the deceptive tasks, the lack of attempted countermeasures (in a first study in which participants were trained to use countermeasures, deception detection accuracy in single participants was 100% without countermeasures, but only 33% with countermeasures (Ganis et al., 2011)), the variability of individual brains and in particular the transferability of the results into real-world situations (Greely & Illes, 2007; Moriarty, 2009).

These concerns are indeed all worth considering, yet one should always bear in mind that other evidence such as witness testimony has not proven to display a high level of accuracy, in particular because the statement itself is already flawed with false memory effects (Eisenberg, 2011) and because ordinary people's ability to distinguish truth from lies rarely rises above chance (Bond & DePaulo, 2006) – which is also true for policemen, prosecutors and judges (Kassin, 2004). Hence, opinions such as Schauer's (2010) should also find consideration:

> ...the admissibility of neural lie-detection evidence must be based on an evaluation of the realistic alternatives within the legal system and not on a non comparative assessment of whether neural lie detection meets the standards that scientists use for scientific purposes (Schauer, 2010, p. 102).

To be perfectly clear, this cannot mean that unsuited lie detection methods should be admissible *because* other unsuited evidence is also admissible. But it cannot be either, that fMRI-based lie detection must meet stricter criteria than other types of evidence. As mentioned, this article is not about the criteria for scientific reliability and validity under the Frye or Daubert standard[25] but about the legal admissibility of an allegedly just good enough neuroscience-based lie detection test. However, the situation in Germany concerning the admissibility of scientific evidence shall be shortly examined.

In German criminal law, the court is allowed to reject a motion to take evidence by the defence only under the regulations of §§ 244, 245 German Criminal Procedure Code (*Strafprozessordnung, StPO*). The *only* criteria for admissibility of a certain scientific method is whether the method is not *"thoroughly unsuited"* (*"völlig ungeeignet"*, § 244 s. 3 ss. 2 var. 4 German Criminal Procedure Code (StPO)).

But when is a scientific method considered not thoroughly unsuited? In contrast to the discussion in the USA, where there are quite a few judgements concerning this issue and where the Supreme Court in its 1993 Daubert decision suggested several non-exclusive factors[26] to consider to aid the trial court in its

[25] For a discussion on the criteria for scientific reliability under the Frye or Daubert standard, see e.g. Greely & Illes (2007) and Moreno (2009).

[26] Such as (1) whether the theory or technique can be tested and has been tested; (2) whether the theory or technique has been subjected to peer review and publication; (3) the known or potential rate of error of the method used and the existence and maintenance of standards controlling the technique's operation; and (4) whether the theory or method has

determination of whether an expert's testimony is suitable, the German judiciary is rather unhelpful and even contradictory in parts. Interestingly, among the few verdicts that concern the question of when a method is not *"thoroughly unsuited"*, are the 1954 and the 1998 *Bundesgerichtshof* (German Federal Court of Justice) rulings about the admissibility of polygraph lie detection. In the earlier verdict, the court said that the suggestion that an untruthful answer leads to an increased reaction of the vegetative system would have to be an *"established well-known fact"*, this would be *"essential"* for the use of scientific methods in court (*Bundesgerichtshof*, 1954, p. 333). Further explanations were not given. The 1998 opinion is much more detailed. However, the court sets quite a high standard: A scientific method (such as the CQT applied with the polygraph) is reliable only if it was a method that was *"generally and free of doubt to be considered as correct and reliable by the relevant experts"* (*Bundesgerichtshof*, 1998, p. 319). This formula now shows great similarity with the 'general acceptance' rule of the *Frye* test from 1923 (the criterion survived as *one* of the factors that should be considered in the *Daubert* standard). The bad news though was that the court neither wasted a single syllable on deducing, explaining or underpinning this standard nor on the question of which criteria would apply for establishing the relevance of the experts at issue. In addition, no references can be found as to different court decisions or relevant literature. It seems as if the standard had appeared from nowhere. Taking into account the wording of § 244 s. 3 ss. 2 var. 4 of the German Criminal Procedure Code (StPO), according to which the evidence must be *"thoroughly unsuited"* for the court to be allowed to reject the respective motion, establishing such a high standard is rather surprising. In addition, considering that the question of the polygraph's reliability was crucial for the judging of its admissibility, the court's silence here is even more surprising.

If the wording of § 244 s. 3 ss. 2 var. 4 of the German Criminal Procedure Code (StPO) is taken seriously though, one has to come to a different conclusion. Although it could surely be asked whether the phrasing was well chosen, as 'unsuited' is not capable of forming a comparative, it can be said however, that the legislature wanted to make a point by putting 'thorough' in front of 'unsuited'. Another decision of the *Bundesgerichtshof* points in the same direction of interpretation. Only if a certain method was *"not thought through"* it could be regarded as *"thoroughly unsuited"*, the court said. But as soon as *"just the slightest inferences"* could be drawn from an expert opinion, the scientific method applied by the expert could no longer be considered *"thoroughly unsuited"* (*Bundesgerichtshof*, 1997, p. 339).

An interpretation of this kind does justice to the wording of § 244 s. 3 ss. 2 var. 4 of the German Criminal Procedure Code (StPO) much more than the strict standard of the 'relevant experts'. The few voices of the literature support this view. For Engels (1981), the unsuitability of the method must be established on the basis of a *"non falsifiable well-known fact"* (Engels, 1981, p. 36). According to Grünwald (1993), it must be *"impossible"* (p. 98) that the method has something to say about the fact in question. For Zwiehoff (2000),

---

been generally accepted by the scientific community.

the fact that there is a controversy in the scientific community over the suitability of a method is insufficient for it to be considered *"thoroughly unsuited"*.

By accepting this much lower standard, it seems that the *Bundesgerichtshof* was wrong when it ruled the CQT inadmissible.[27] Even when considering the many submitted studies that have all, with a few exceptions, shown rates of at least 70% accuracy for the CQT (for an overview, see Honts, 2004), it might well be justified to say that the method's suitability is not very good, because of the many possible objections such as for example the ones expressed by Greely and Illes (2007) mentioned above. But it is unlikely that such neuroscience-based lie detection tests are not at least a bit better than chance in determining the tested person's truthfulness (see Seiterle, 2010). It has to be emphasised once more that – at least for the use of neuroscience-based lie detection as a means of *exculpation* – it is not at all necessary for the method to have a very high validity or to even be perfect – this aspect seems to have a tendency of being neglected in the discussion. In particular under German law, it suffices for the method not to be *"thoroughly unsuited",* that it was only *one percent* better than chance in assessing the accused's truthfulness.

There are not that many fMRI studies yet that aim at determining veracity on an *individual* level, so that it might still be too early to call any fMRI lie detector better than completely unsuited. It is not unlikely that the moment will come sooner than expected though, that it must be regarded as better than flipping a coin when an accused's (or witness') veracity is to be judged, as is already the case with polygraph lie detection (see also Schauer, 2010).

Legal admissibility of a suitable non-invasive fMRI-based lie detection method

Coerced use prohibited – procedural 'nature' of the neural 'statement'

§ 136a of the German Criminal Procedure Code (StPO) prohibits the use of certain interrogation methods such as torture, other physical harm, deception, etc. and is the statutory expression of the fundamentally guaranteed right not to incriminate oneself – the self-incrimination clause (in the US constitution protected under the 5th Amendment).

It is already doubtful whether a coerced use of fMRI lie detection is feasible at all. For a coerced use, it would be necessary to fix the subject on the table and force him to listen to or watch the examiner's commands and make sure that he actually understands their meaning.

But given that coerced use was possible for the purpose of argumentation, at first sight it seems that treating an accused this way would certainly fall under § 136a of the German Criminal Procedure Code

[27] This view is shared by legal scholars (see Putzke, Scheinfeld, Klein, & Undeutsch (2009) and Scheffler (2008)) as well as by psychophysiologists (Offe & Offe, 2004; Fabian & Stadler, 2000).

(StPO). However, this conclusion might be rash, for what the accused does actually 'utter' during the lie detection test differs from what is traditionally protected by the self-incrimination clause (§ 136a of the German Criminal Procedure Code (StPO)): his *verbal* testimony. On the other hand, the self-incrimination clause does not prevent the accused from being the source of *physical evidence*: § 81a of the German Criminal Procedure Code (StPO) permits the examination of the accused even against his expressed will. Physical tests like blood tests, fingerprints, etc. are permitted to be carried out according to §§ 81f of the German Criminal Procedure Code (StPO). fMRI lie detection has at least elements of a physical examination, too. It is evident that the traditional distinction between *physical* evidence that can be retrieved without the suspects's consent on the one hand and *testimonial* evidence that is protected by the right not to incriminate oneself on the other hand, does not provide clear guidance in this matter. It has already been proposed to think of something like a third way to solve this problem (Thompson, 2007). However, as long as legislature remains silent, it is the jurists who have to decide whether the 'statements' made during a forced neural lie detection test fall within the scope of § 136a or § 81a of the German Criminal Procedure Code (StPO); at least for German law that means: *tertium non datur*.

As there are no precedents that deal with this issue, one can only try to find arguments. In German criminal procedure law, the classic distinction is as follows: The state must not compel the accused into doing something *actively*. As long as this is not the case, the accused might be 'used' as physical evidence. Obviously, this line of distinction becomes difficult when deciding about the lawfulness of a (if ever feasible) coerced neural lie detection test: The accused may be fixed and forced into an fMRI scanner, but he would not be forced to do anything *active*, for his brain activity 'happens' involuntarily (Thompson, 2007). Stoller and Wolpe (2007) sum up the problem as follows: *"For the first time in human history, the state may be able to obtain information directly from the brain against a suspect's will"* (Stoller & Wolpe, 2007, p. 367; see also Fox, 2009).

In the German discussion, it has been suggested to use the question whether there is *communication* as crucial criterion for the distinction between physical and testimonial evidence (Groth, 2003). Whenever there is communication between the accused and the criminal prosecution authorities – or the examiner as their representative – it would concern testimonial evidence. However, this criterion turns out not to be the most useful tool. Firstly, it is disputable what exactly is understood by the concept of 'communication'. Secondly, it is not always possible to achieve precise results. How for instance would the case be judged that – one day – a computer programme developed the questions for our neural lie detection test, a computer-generated voice read them out automatically to the accused and the measures would be analysed by special software? It does not seem easy to subsume this course of events to 'communication' between two parties. Or the other way round: Groth (2003), who suggests the communication criterion, goes with Watzlawick and colleagues (1985), who state that two people *"cannot not communicate"* (Watzlawick, Beavin, & Jackson, 1985, p. 51). Why would taking a blood sample from an accused then not fall under 'communication', as the accused sends information (incorporated in his blood) to the doctor/nurse, which influences her and causes her reaction? No communication?

Beyond this criticism, it is not made clear *why* the *formal* fact that there is communication going on should be relevant to our problem. Unless – something not done by Groth (2003) – it is *explained* why this

criterion should be more than a mere phenomenological observation, it cannot be used as an argument in a satisfactory way. What is needed instead is a *material* criterion. Here, the most convincing approach lies in asking whether the suspect is to reveal *knowledge* related to the criminal act.[28] If this is the case, the use of force to retrieve information is prohibited. This is also the main reason the privilege against self-incrimination was installed in the first place: Other than in the infamous inquisition process, the self-incrimination clause attempted to guarantee the accused's right of disposal over his knowledge. The accused has the right to decide if and to what extent he is willing to reveal his knowledge related to the criminal act to the investigators. This said, it is not important *how* the statement is being made, whether verbal or non-verbal, consciously or unconsciously. Physical evidence like blood samples, on the contrary, do not concern the accused's knowledge related to the criminal act, which is why it is allowed to retrieve it against the subject's will.[29]

Accepting the knowledge criterion does not only produce rather clear results, it also provides a *reason* for the different legal treatment of physical evidence on the one hand and testimonial evidence on the other hand. It lies in the fact that traditionally the mind of man is regarded as more valuable than his mere physical being – which is not surprising given that the mind is what makes man unique and distinguishes him from animals. Human dignity, autonomy and the subjectivity of man are not determined by his biological existence but by his special ability to mentally and emotionally make and exercise his will (Verrel, 2001).

When applying the knowledge criterion to any kind of lie detection, the outcome is evident: It is the *principle* of lie detection in the setting of criminal procedure that the suspect – at least indirectly via interpretation by the examiner – reveals knowledge (or the opposite) related to the criminal act. In the case of the CQT, the suspect is asked whether he took part in the crime and in the case of the GKT, the aim of the examination is already been mentioned in the name of the questioning technique. As a result, the accused must not be compelled to reveal his knowledge about the criminal act in a neural lie detection test, because he is protected by the self-incrimination clause.

Use without the accused's consent – use in secret

By accepting *knowledge* related to the crime as being the decisive criterion for distinguishing between testimonial and physical evidence, another problem can be addressed that has not yet been much

[28] See the discussion in the USA, where also a distinction is made between the act of communicating (*disclosure* of the contents of one's mind) and the product of this communication (the *contents of one's mind*), described in some detail by Stoller & Wolpe (2007); see also Thompson (2006). For Germany, see Frister (1994) and Verrel (2001).

[29] See Schmerber vs. California (1966), where the Supreme Court ruled that the police could compel a suspect to provide a blood sample for analysis in order to determine whether he was intoxicated, because this process was in no way testimonial (Stoller & Wolpe, 2007).

discussed. There are attempts to develop mind-reading devices that work remotely, by using Near-Infrared Spectroscopy (see Greely & Illes, 2007), Infrared Photography (Pavlidis, Eberhardt, & Levine, 2002) or similar technologies. Under the further prerequisite that no questioning technique would be necessary for that kind of lie detector, the following scenario could become reality one day: The accused would be filmed while making his statement or even if he chooses to remain silent. An expert witness would then analyse the data and give a report on the 'content' of the subject's mind and/or on the veracity of his statement. No coercion whatsoever would be needed to retrieve knowledge linked to the criminal act. However, as has been shown, the self-incrimination clause (§ 136a of the German Criminal Procedure Code (StPO)) would prohibit the use of such a device as it gives the accused the right to dismiss statements of whatever kind that concern his knowledge about the criminal act in question.

Use with the accused's consent

It is much more difficult to determine whether the use of a fMRI-based lie-detection technique should be admissible if the accused consents to it or even demands it for the purpose of exoneration.

Would the accused's consent be voluntary?

Even if a neuroscience-based lie detection test was generally admissible in criminal courts with the accused's consent, it is yet to ask whether that consent could be acknowledged as voluntary at all. The German Supreme Court (*Bundesverfassungsgericht*) held in 1981 that a polygraph test would infringe the accused's personality rights because his consent would *per se* be involuntary (*Bundesverfassungsgericht*, 1982, p. 375). Protection against state acts was only unnecessary if the individual made a real choice, the court stated. But according to the *Bundesverfassungsgericht*, this freedom of choice was not given to an accused for who – depending on the evidence at hand against her – conviction was the more or less secure outcome when *rejecting* the polygraph test. In this situation, consenting to a lie detection test was an option that the accused could not reasonably refuse. Although the Supreme Court's view was barely agreed with, it cannot be completely ignored[30], for it may be agreed with the court that this is not the classic situation of a free choice. It must be considered that the accused here offers his basic rights, like his human dignity or his common personal right, to be at least touched by the government-ordered neuroscience-based lie detection test *solely* because he wants to prevent the state to interfere with *other* rights like his right to freedom (in case of a prison sentence) or his right to property (in case of a fine). This indeed can be called a dilemma as it seems impossible for the accused to protect all his rights at the same time and he is forced to sacrifice at least one of them. The accused finds himself in an emergency, in which at least if innocent, he must decide

[30] It is important to note here that this verdict does not create a precedent.

in favour of the lie detection test if he wants to keep the possibility to achieve acquittal – in *this* respect one might follow the Supreme Court's argumentation.

What the court does not see however, is that a lack of freedom of choice must not necessarily lead to the consent becoming invalid. In many parallel cases (such as §§ 56c s. 3 § 183 s. 3 German Criminal Code (*Strafgesetzbuch, StGB*): consent to medical treatment, although the convict only consents in order to avoid enforcement of the prison sentence) consent to the infringement of her rights is acknowledged as valid, although it is no more than *"semi-voluntary"* (Amelung, 1999, p. 384). According to Amelung (1999), the consent in these situations might be called *"encroachment-relaxing consent"* (Amelung, 1981, p. 105). The basic idea behind the aforementioned regulations is the principle of proportionality: State organs have to choose the most lenient way when interfering with the individual's sphere. If the individual faced such interference it could be the most lenient way to let him *participate.* It has to be left to the person affected, which of two (at least formally) legal detriments to his rights he wants to accept, because he is the only one who can judge which of the potential losses affects him less. To summarise, the consent in these cases is acknowledged as valid despite the predicament, (only) because with his consent the accused is able to mitigate detriments that otherwise would be disproportionate. Lie detection aiming at exoneration is not about the accused reducing the extent of a *legal* detriment by his consent altogether. The accused rather wants to *prevent* that he has to take – in case of his innocence – an *illegitimate* detriment of his basic rights. That is why it is not about the 'classic' case of an *"encroachment-relaxing consent"* which could be explained through the principle of proportionality. Hence, a different justification is required if this *"encroachment-preventing consent"* should be acknowledged as autonomous despite the obvious crises the accused is in.

The true reason why the accused's consent in a lie detection test cannot be declared insignificant is as follows: A main goal of the criminal procedure is realising the 'guilt principle': in the inquisitorial system, the court always has to endeavour to *"do everything so that the guilty will be punished according to his guilt and that the innocent will be released from the procedure or acquitted"* (*Bundesverfassungsgericht*, 1987, p. 2663). The court is therefore obliged to investigate the facts, the material truth, so that it is able to make a just decision. But the accused also has to be protected from unjustified burden; if he consents to the state-caused violation of his rights, because he wants to convince the court of his innocence, this wish has to be respected.

From what has been said, it is possible to deduce the further requirements for a consent to be voluntary in this sense: The state is only permitted to intrude in the accused's sphere because of his consent if the aforementioned goals are really served in this way. Therefore, the tool that is used for the exoneration purpose must be a suitable and (otherwise legally) admissible means for the search for the truth. If, for example, the *Bundesgerichtshof's* opinion concerning the reliability of the CQT was correct in that it could not contribute in the slightest to establish the truth, this test would already be inadmissible because the accused's consent could not count as voluntary. The same conclusion would result, if there were infringements that would not serve the search for the truth but that the accused would offer in *exchange* for acquittal (if he for example offered to go to a mental hospital for a certain amount of time). If in contrast, the consent's validity was derived solely from the idea of the accused's autonomy, like it has been done (Brandis 2001), one would be forced to acknowledge also senseless 'sacrifices'. This argumentation overlooks that it

is about state encroachments, which without a statutory authority can *only* be justified by the accused's consent. State organs are not, however, entitled to harm citizens just upon their wish as soon as they are in *any kind* of crisis, although this would be the consequence of a justification that was arguing with the idea of the individual's autonomy.

In the case of fMRI lie detection, the just described requirements for 'voluntary' consent are met: It is assumed here that our lie detector is reliable at least to a certain extent and it would therefore be a suitable means for the search for truth. The accused therefore would be able to avoid being convicted because of wrong assumptions to a – in case of his innocence – illegitimate punishment just through partially sacrificing his personality rights. As long as there was no additional pressure, the consent would be 'voluntary' or at least valid.

Would the use of neural lie detection violate the subject's personal rights/human dignity – despite her valid consent?

Having acknowledged the consent of the accused who wishes to exonerate himself as voluntary, it has to be clarified whether law *permits* him to consent to a state measure that potentially affects his personality rights or even his human dignity. The *Bundesgerichtshof* (German Federal Court of Justice*)* ruled in 1954 that an accused must not consent to a polygraph test. The court argued – although it has to be pointed out that there is not much of a detailed argumentation to be found in the reasons for the judgement – that a polygraph test ordered by the court does violate the accused's human dignity, even if he consented to it. It seems obvious that this decision was influenced by the post-"Third Reich"-atmosphere, where there was a general scepticism towards any method that had to do with – alleged – interference with a subject's psyche. Despite the fundamental lack of argumentative depth, German criminal law literature supported this verdict almost unanimously for at least 20 years. And yet it is not an absurd question to ask whether the *Bundesgerichtshof's* judgement could be reasonably justified. It has to be scrutinised whether basic rights might be turned against the bearer of the rights. In general, it is about the right to disposal of the subject's basic rights. Under German law, it is acknowledged that the common personality right (art. 2 s. 1 in conjunction with art. 1 s. 1 German Constitution) is a disposable right. Hence, the infringement of this right by the state use of an fMRI lie detector would be justified by the subject's informed consent.

This result is not as clear with the case of human dignity (art. 1 s. 1 German Constitution). As mentioned, the *Bundesgerichtshof* did not accept the consent as justificatory in its 1954 decision. However, after a 20-year-discussion that started off in the late 1970's and that focused on the *innocent* accused whose last chance of producing evidence would be to take part in a suitable method of lie detection (see Schwabe, 1979), in a second verdict in 1998, the *Bundesgerichtshof* changed its mind. If the accused voluntarily consents to the lie detection test, there would not be any violation of his human dignity, the court said (*Bundesgerichtshof*, 1998, p. 317).

Case closed? Rather not. There is another passage in the opinion for the judgement that reads differently: In order to establish a lie detection test with the consent of the accused as non-dignity-violating, according to the court it was *"crucial"* that it was impossible that the method in question could not provide a *"specific lie reaction"* (which was not the case with the CQT/polygraph, the court said) (*Bundesgerichtshof*

1998, p. 315). The court ties in the 1954 verdict here, according to which it was possible with a polygraph test to look into the *"soul"* of the accused (*Bundesgerichtshof*, 1954, p. 335). If with a test one could indicate a *"specific lie reaction"*, according to the 1998 verdict, this would be equivalent to looking *"into the soul"* (*Bundesgerichtshof*, 1998, p. 315). So, for existing methods, the *Bundesgerichtshof* ruled that the consent is sufficient to justify the implicit infringement of the accused's human dignity. But as soon as there was a *"specific lie response"*, the lie detection test would be inadmissible due to a violation of dignity. At least it seems that this is what the court is saying, although it did not make its point totally clear and there is some debate amongst the criminal law experts as to whether this was really what the court *wanted* to say (if it tried to say something here, not even that seems certain). So the crucial question is still, whether the accused's will is of such importance that it could justify a state order (like ordering and performing a lie detection test) that would violate his human dignity if it were done without or even against his will. This issue concerns the disposability of the guarantee of the constitution to protect every human's dignity.

In the outcome, the *Bundesgerichtshof* was right in its 1998 verdict. In a 'free democratic basic order' as Germany has, the guarantee for human dignity has to be understood first and foremost as recognition of the individual's right to self-determination. Religious beliefs like they might have underlain the 1954 decision about a lie detection test looking *"into the soul"* of the accused cannot – in a secular and pluralistic society – be used against personal autonomy. Against this background, any attempts to justify 'hard paternalistic' approaches to compel someone into behaviour that at least in a way was for his own good, is destined for failure.

There are other verdicts amongst German judiciary that come to a different conclusion. According to these, human dignity is an *"objective, indispensable"* value, *"of which recognition the individual must not renounce"*, because its importance goes *"beyond the individual"* (*Bundesverwaltungsgericht* (Supreme Administrative Court), 1981, p. 279). However, in no case, the respective court gave the slightest reasons for its decision, which is why it is impossible to further comment on these singular verdicts, even though they were pronounced by federal courts.

Others try to justify the prohibition of a certain – voluntary – conduct in a different way: The respective conduct was to be prohibited because permitting this kind of conduct would have the potential to severely change the idea of humankind. Schmitt Glaeser (2000), for instance, has attempted to give reasons for the prohibition of TV shows like "Big Brother". According to him, if a certain conduct undermined *"the infrastructure of human dignity, understood as legal and factual substructure of its recognition, respect, its protection and its development"* (Schmitt Glaeser, 2000, p. 400) and therefore endangered or destroyed the condition for a dignified life of many or all human beings, this conduct should be prohibited, even and particularly if it were with the actor's valid consent. As it is evident that voluntary lie detection, even if done with neuroimaging techniques, does however probably not have such strong impacts, it does not seem necessary to elaborate on this issue. But even if one assumed that lie detection – perhaps some futuristic kind of lie detection we are unable to even imagine yet – had that potential, the argument of the idea of humankind was wrongly placed within the scope of human dignity. The reason for this is that when talking about a change in the idea of humankind, one inquires about the *consequences* of the admission of a certain conduct. This has nothing to do with the specific subject of human dignity, but belongs to the general aspect

of the possibly overriding interests of the general public or third persons (see below).

As a consequence, there is no voluntary conduct that could be prohibited under the aspect of the allegedly violated individual's human dignity, whether it is taking part as a dancer in a *"peep show"* (*Bundesverwaltungsgericht*, 1981, p. 274) or in an 'event' that has become famous under the term 'dwarf throwing' (*Verwaltungsgericht* (Administrative Court) Neustadt, 1993, p. 99). Not even 'voluntary' torture, if there was a logically satisfying concept of it, would violate the suspect's human dignity. This does *not* mean, however, that conduct like the aforementioned would necessarily be permitted. In some cases, especially for voluntary torture with the aim of strengthening one's credibility in court, it seems quite facile to find different reasons for it to be prohibited – but the consenter's human dignity does not seem to be a valid one.

Coming back to the *Bundesgerichtshof's* 1998 judgement, it turns out that it would not be practicable to justify the inadmissibility of a lie detection method used with the accused's informed consent, that provided a *"specific lie response"* or worse, with which it was possible to *"look into the soul of the accused"*, with the argument that it violated the accused's human dignity (Spranger, 2009). If it was the court's intention to keep the door open for falling back on the metaphysical view of the 1954 judgement – which cannot be decided from the wording of the opinion (see above) – and thus to prepare for the prohibition of a suitable neuroscience-based lie detector once it is there, this attempt has to be rejected. (Naturally, the same applies to those opinions that follow from the 1998 decision – and approve of – the 'final end' of any kind of lie detection in German criminal courts (Kühne, 2010)).

Prevailing interests of the general public/third persons

Interests of the general public

As mentioned above, a possible objection could be that the concept of the idea of man could be affected by the widespread use of fMRI methods in criminal courts. But in this case one would have to show that the admission could have such severe consequences, that these would have to be avoided even at the expense of the – sometimes vital – interests of the accused in question. Quite rightly nobody has ever tried to use this kind of argumentation against lie detection yet, considering that lie detection methods are used in many – also western – countries. The fact that it was not the polygraph but a brain imaging method makes no difference as fMRI, EEG and the like are also generally accepted and widely used for the most diverse purposes. The public charge argument does not become valid unless a perfect brain-reading machine is developed that could be purchased and used by almost everybody.

Interests of third persons/indirect pressure upon the future accused

Much weightier appear the concerns about the interests of people who may be accused of a crime in the future. What if a suspect does not utter the wish to be tested in order to prove his innocence? Would such conduct not be interpreted as an admission of his guilt by the judge(s)/the jury? And would this not impose an (indirect) pressure upon the accused to consent to the use of the lie detection method against his original intention? In the German discussion about the admissibility of lie detection in criminal court, for some authors only this argument of 'indirect pressure' remains (Frister, 1993). Some reject the admissibility solely on the basis of the consequences of 'indirect pressure' which they call *"almost impossible to calculate"*

(Rogall, 2010, para. 95).

Often neglected in this context is the fact that it is not about the alleged fact of implicit pressure on the accused. This pressure can only be relevant for the question of admissibility of lie detection if it (also) had *legal* consequences. But before these legal consequences are considered, the right not to incriminate oneself might be at issue. This would have to be scrutinised as by definition it has to be ascertained whether lie detection in court would lead to a significant pressure on the accused *at all*. In order to show this, several assumptions have to be made. Some of these assumptions are no more than speculations about mental processes within judges/jurors, whereas others can be supported by empirical research.

Firstly, judges/jurors would have to interpret a failure to undertake the test by the subject as circumstantial evidence for his involvement. For this purpose, it is useful to compare this to the discussion about the right to silence and the question if and to which extent it is permitted to draw negative inferences from the fact that the accused does not reply to the accusations. For a long time, there was not much doubt that silence meant the accused had something to hide, which usually meant: his guilt (Bentham, 1962). Only later it became accepted that silence might have many others reasons – general fear or intimidation, protecting others, political attitude, etc. – and that at least *solely* from the accused's silence the court/jury must not deduce his guilt.[31] When taking a closer look at the situation in which the accused refuses to undertake a lie detection test though, you will find important differences to the case of the accused remaining silent. On the one hand, it has to be considered that we are dealing with a method that classifies 'non-guilty' subjects as 'guilty' with a significant probability, which means that over every truly innocent accused hangs the sword of Damocles of receiving a wrong incriminating test result – a 'false positive'. This is a good reason for every subject not to ask for the test! This reasoning would also be taken into consideration by the court/the jury and consequently a lack of desire to be tested would not be interpreted as a sign of guilt.

However, this conclusion would be premature, for one has to distinguish between different situations in a criminal trial. If basically all the evidence pointed to the accused and if a subject was just about to be convicted, a false positive neuroscience-based lie detection test result would not deter most of them from consenting to be tested. In this case, judges/juries would probably be even more tempted to draw negative inferences from a suspect who would *not* take his last opportunity to prove his innocence in this situation. But it gets even more complicated: When conviction is almost inevitable, the court/jury is not in need of an additional piece of evidence, so that the 'evidence' of a lack of a consent to undergo a lie detection test plays no role in the first place. This leads to the assumption that in this situation an accused would feel no (additional) pressure to consent to a lie detection test against his true intention, because he does not have to

[31] In England and Wales (and also Northern Ireland, Criminal Evidence Order 1988), it is at least under certain circumstances allowed for the judge/jury to draw inferences that appear 'proper' (sections 34f. Criminal Justice and Public Order Act; 1994).

fear that this decision would be regarded as suspicious.

But what about a case in which there is some evidence against the subject, but just not enough to convict him, in which in other words, one final piece of the puzzle might still be missing? This lacking piece of evidence could now be delivered by exactly the negative inference a jury/judges might draw from the fact that the accused refuses to 'prove' his innocence by means of a suitable neuroscience-based lie detection test. But this consideration is not yet sufficient: Subjects would *only* feel pressured to offer this kind of evidence, if it was about a lie detection method that had a very high specificity, i. e. a test that has a very low rate of classifying non-guilty suspects as 'guilty' (a low false positive rate). If a test's false positive rate was high, every accused, and especially an innocent suspect, would have very good reason to refrain from taking part in a neural lie detection test. As already mentioned above, this reason would apply to every subject and it is evident that – we are still speculating about complex inter-psychological processes – the judge(s)/jury would not conclude that this the lacking piece of evidence. As there would be no inferences, the accused would feel no pressure to do what he did not originally intend to do.

Here is a first result: The problem of the 'indirect pressure' on subjects that might be caused by admitting lie detection in court is very complex, but turns out not to be as incalculable as it is widely feared. A possible violation of the right not to incriminate oneself might *only* occur if there is some evidence against the accused but perhaps not enough to convict him without further pieces of evidence, and if it concerns an almost perfect neuroscience-based lie detection test, a test that has a very high specificity. In any other case, the problem of 'indirect pressure'/the self-incrimination clause would not emerge at all.

And even for the remaining cases the discussion is not yet over. At least one further requirement must be met: In accordance with the right to remain silent, inferring guilt from the fact that the subject refrains from requesting a lie detection test would be subject to the prohibition of exploitation: Even if it might appear just reasonable to interpret such a conduct as admission of guilt, the judges/jurors are legally prohibited to use it as evidence against the accused, because they would thereby violate the accused's right not to incriminate himself (*Bundesgerichtshof*, 1998; Beck, 2006).

The interesting question now is whether in the real world the decision-makers would be *capable* of complying with that prohibition. One can easily find pros and cons on a once again mere speculative level, but there are also a few recent studies that investigate this question empirically (Wistrich, Guthrie, & Rachlinski, 2005; Kassin & Sommers, 1997). The results are inconsistent, but there are some intriguing findings. It seems that it is harder for judges to ignore information like previous convictions, conversations between the accused and his counsel, or details from confidential amicable agreements than information that was gained by violating the accused's guaranteed right to counsel. When considering that an accused, who claims to be innocent and faces a certain danger to be convicted, shows no interest in a – sometimes even literally life-saving (USA) – very promising way of exonerating himself from the accusations, might have something to hide, namely his guilt, it seems likely that judges/jurors do not always have the capacity to fully ignore this 'evidence'.

If this is accepted it is only now that the question arises which *rights* of an accused might be affected in the above-mentioned case, in which the conviction is yet unsure and the available 'evidence' of the lack of a desire to be tested might be the last piece of the puzzle. Here one is forced to differentiate once again: The

right to testify is violated if the suspect has not made the decision whether he would like to go for the test yet, because due to the indirect pressure the choice whether or not to undergo the lie detection test is not free anymore. Those subjects who actually decide in *favour* of the test, although this decision contradicts their original attitude, suffer an infringement of their personality rights, because the consent to a method that has the potential to affect this right cannot be regarded as voluntary anymore, so that the consent loses its justificatory power. Finally, admitting fMRI-based lie detection would infringe the right not to incriminate oneself of those who do *resist* the pressure and refrain from consenting to a lie detection test even in this precarious case, because this – permitted! – conduct would (probably) be interpreted as admission of guilt and the accused would at least indirectly be *"compelled…to be a witness against himself"* (US constitution, 5th Amendment). Only for this – obviously very limited – case the doubts expressed by quite a few authors are justified. But of course the concerns are only justified if it is made sure that no *additional* pressure is put upon the subject, neither by the police or the court nor by the prosecution.

But even if it was established that – at least in certain cases – admitting fMRI lie detection would lead to an inevitable violation of the self-incrimination clause for some accused, this does not necessarily mean that their rights would come first. It has to be clarified whether the interests of an accused to have his procedural rights protected and preserved as thoroughly as possible take priority, or whether the interests of an accused to have access to a means that gives him a promising chance of proving his innocence take priority. Once again, the consequence is not consistent. For instance, if we are dealing with a situation where the case against the accused is weak, so that conviction is more or less out of the question, the interests of the accused to gain just an additional possibility to undermine his innocence have to take second place to the interest of an individual who may be accused of a crime in the future to complete protection of his rights. However, in the case that sparked off the 'modern' discussion about the admissibility of polygraph lie detection in Germany in the late 1970's, the outcome is different: When a person who is accused of a crime, in particular those who are innocent, according to the body of evidence has to seriously fear conviction – and there is no other evidence for the defence – he has a *vital* interest in being given the promising opportunity to produce evidence of his innocence. There cannot be much doubt that the interest in not to be – in the case of innocence unlawfully – given a prison sentence or, worst case (USA) to be sentenced to death, overrides the interests of people who may be accused of a crime in the future.[32]

To summarise: The motion to take evidence by way of consenting to a neural lie detection test, which would at least indirectly allow conclusions to be drawn about the veracity of the accused's statements, is admissible without further problem in most cases. In particular, this is the case when it concerns a lie detection technique that is *not* perfect in the sense that the probability for false positives is significantly high. It has to be emphasised that as long as fMRI or any other form of lie detection does not show a very good

[32] This view is shared, for example, by Amelung (1982) and Schwabe (1979).

specificity, rights and interests of the accused will (probably) not be affected and the test is admissible in total. This means, that even if the testimonial nature of the physiological 'expressions' is accepted, the voluntary use of fMRI-based lie detection is admissible under German law if the method is suitable (see above). Only a method with a very high specificity could cause problems under certain aspects, mainly a possible impairment of the right not to incriminate oneself, because only then a suspect is likely to feel an indirect pressure to go for the neural lie detection against his initial intention. But even in this case this conclusion is only valid if one accepts that judges/jurors are not (always) capable of ignoring the prohibition not to exploit the 'evidence' that a truly innocent accused would certainly apply for a lie detection test and that therefore the fact that a subject refrains from doing so must be an implicit admittance of his guilt. In addition, this conclusion is only valid if there is most likely not enough evidence against the accused, as otherwise the interest of an accused to take advantage of a reliable way to exonerate himself from the accusation and thus prevent unlawful conviction would override the interest of an individual who may be accused of a crime in the future not to accept violation of his rights, in particular his right not to incriminate himself[33].

However, all these considerations can only claim to be correct if the subject's voluntariness is secured. As soon as there was the slightest pressure on the subject to consent to a lie detection test against his original intention *beyond* the mentioned implicit pressure, there would be a different case and many concerns would gain importance again. In order to avoid this, it would be important to prohibit the police or the prosecution from even talking about the possibility of a lie detection test, let alone suggest it, even for the purpose of exoneration. The test should only be admissible if the accused takes the initiative without having been influenced in any way by state authorities. Only if this requirement is met, the admissibility of fMRI-based lie detection seems rather unproblematic.

Conclusion

In comparison to the US legal system, the standards for the admissibility of scientific evidence are lower under the rules of the German Code of Criminal Procedure (StPO). Accordingly, a scientific method must be *"thoroughly unsuited"* for it to be regarded as inadmissible by the court. Hence, a neuroscience-based lie detection method like fMRI could in theory be considered admissible in Germany sooner than in the US, although in reality this seems rather unlikely, as the *Bundesgerichtshof's* (German Federal Court of Justice) attitude toward the accuracy of physiological lie detection methods (like the polygraph) is very

[33] For a more detailed description of the rights of an accused who wishes to use a lie detection method for the purpose of exculpation and the – potentially conflicting – interests of different accused not to be violated in their fundamental procedural rights, see Seiterle (2010).

sceptical.

Given the suitability of neural lie detection, *purely* legal issues have to be raised. Firstly, the question whether a compelled (or secret) use would be legal depends on how you assess the nature of the subject's brain activity that is measured during the test. There are good reasons for regarding brain activity as *testimonial*, for which reason the use of neuroscience-based lie detection would fall under the self-incrimination clause (5th Amendment in US law). As a result, the secret or forced use would be prohibited and only lie detection with the accused's *consent* is worth further examination. Despite the crisis a suspect finds himself in when facing conviction, his consent would have to be acknowledged as at least *"semi-voluntary"* and therefore be valid.

In contrast to the *Bundesgerichtshof's* 1954 ruling and several legal scholars, the accused's human dignity cannot be turned against him: There is no – at least no *legal* – duty to behave with dignity. As long as consent is valid, the actual testing will therefore not infringe upon the accused's human dignity or interfere with any other of his rights.

It is prohibited for the judge/jury to draw negative inferences from the fact that an accused does not consent to undergo a potentially exculpating lie detection test. Hence, the rights of the accused could only count as argument against the admissibility of lie detection in criminal court, if one assumed that judges/jury are not fully capable of complying with this prohibition. In this case, the rights of the accused against self-incrimination would potentially be threatened. But only within limits: In the case where the accused is expecting acquittal, the interests of individuals who may be accused of a crime in the future prevail over the interest of the accused at issue to exonerate himself by means of a suitable lie detection test. In every other case suitable lie detection for the purpose of exculpation is legally admissible in German criminal courts.

Finally, the analysis leads to the assumption that for the (purely) legal issues, the fact that not the polygraph but fMRI or another neuroscience-based lie detection method is used, is not crucial (Spranger, 2009; Schneider, 2010), neither for the question whether the use of the lie detector with the consent of the accused violates her human dignity, nor for the discussion about the legal 'nature' of the measured physical 'statements' (vegetative symptoms/brain activity), nor for the assessment of the possibly infringed rights of the general public. The most important difference, as far as technical issues are concerned, is that with neuroscience it seems possible that lie detection might be used without the subject even being aware of it. But this only means that another situation has to be legally considered, it does not mean a fundamental difference in the assessment itself.

References

Abe, N., Suzuki, M., Tsukiura T., Mori, E., Yamaguchi, K., Itoh, M., & Fujii, T. (2006). Dissociable roles of prefrontal and anterior cingulate cortices in deception. *Cerebral Cortex, 16*, 192-199.

Abe, N., Suzuki, M., Mori, E., Itoh, M., & Fujii, T. (2007). Deceiving others: Distinct neural responses of the prefrontal cortex and amygdala in simple fabrication and deception with social interactions. *Journal of Cognitive Neuroscience, 19*, 287-295.

Abe, N., Fujii, T., Hirayama, K., Takeda, A., Hosokai, Y., Ishioka, I., Nishio, Y., Suzuki, K., Itoyama, Y., Takahashi, S., Fukuda, H., & Mori, E. (2009). Do parkinsonian patients have trouble telling lies? The neurobiological basis of deceptive behaviour. *Brain, 132,* 1386-1395.

Amelung, K. (1999). Anmerkung zu BGH, Urteil v. 17.12.1998 – 1 StR 156/98 (BGHSt. 44, 308). *Juristische Rundschau*, 382-385.

Amelung, K. (1982). Anmerkung zu BVerfG, Beschl. v. 18.8.1981 – 2 BvR 166/81. *Neue Zeitschrift für Strafrecht*, 38-40.

Amelung, K. (1981). *Die Einwilligung in die Beeinträchtigung eines Grundrechtsgutes: Eine Untersuchung im Grenzbereich von Grundrechts- und Strafrechtsdogmatik*. Berlin: Duncker & Humblot.

Beck, S. (2006). Unterstützung der Strafermittlung in den Neurowissenschaften? – Einsatz von Verfahren funktioneller Bildgebung als „Lügendetektoren" im Strafprozess. *Juristische Rundschau*, 146-150.

Bhatt, S., Mbwana, J., Adeyemo, A., Sawyer, A., Hailu, A., & VanMeter, J. (2009). Lying about facial recognition: An fMRI study. *Brain and Cognition, 69*, 382-390.

Bond, C., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*, 214-234.

Brandis, T. (2001). *Beweisverbote als Belastungsverbote aus Sicht des Beschuldigten?* Frankfurt/Main: Peter Lang.

Bundesgerichtshof (1997). *Strafverteidiger*, 339.

Bundesgerichtshof (1954). *BGHSt*, 5, 333-335.

Bundesgerichtshof (1998). *BGHSt*, 44, 308-328.

Bundesverfassungsgericht (1982). *Neue Juristische Wochenschrift*, 375.

Bundesverfassungsgericht (1987). *Neue Juristische Wochenschrift*, 2663.

Bundesverwaltungsgericht (1981). *BVerwGE*, 64, 279.

Christ, S.E., Van Essen, D.C., Watson, J.M., Brubaker, L.E., & McDermott, K.B. (2009). The contributions of prefrontal cortex and executive control to deception: evidence from activation likelihood estimate meta-analyses. *Cerebral Cortex, 19*, 1557-1566.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D.G., Acharyya, M., Loughead, J.W., Gur, R. C., & Langleben, D.D. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage, 28 (3)*, 663-668.

Eisenberg, U. (2011). *Beweisrecht der StPO. Spezialkommentar* (7th ed.). München: C.H. Beck.

Engels, D. (1981). Beweisantizipationsverbot und Beweiserhebungsumfang im Strafprozeß. *Goltdammer's Archiv*, 21-36.

Fabian, T., & Stadler, M. (2000). Polygraphietest im Ermittlungsverfahren. Möglichkeiten der physiopsychologischen Verdachtsabklärung im Ermittlungsverfahren. *Kriminalistik, 54*, 607-612.

Ford, E. B. (2006). Lie detection: historical, neuropsychiatric and legal dimensions. *International Journal of Law and Psychiatry, 29*, 159-177.

Fox, D. (2009). The right to silence as protecting mental control. *Akron Law Review, 42 (3),* 763-801.

Frister, H. (1994). Der Lügendetektor – Zulässiger Sachbeweis oder unzulässige Vernehmungsmethode?. *Zeitschrift für die gesamte Strafrechtswissenschaft, 106* , 303-331.

Ganis, G., Kosslyn, S. M., Stose, S. Thompson, W.L., & Yurgelun-Todd, D.A. (2003). Neural correlates of different types of deception: an fMRI investigation, *Cerebral Cortex, 13*, 830-836.

Ganis, G., Rosenfeld, J.P., Meixner, J., Kievit, R.A., & Schendan, H.E. (2011). Lying in the scanner: Covert countermeasures disrupt deception detection, *NeuroImage, 55*, 312-319.

Greely, H. T., & Illes, J. (2007). Neuroscience-based lie detection: The urgent need for regulation. *American Journal of Law & Medicine, 33*, 377-431.

Groth, K. (2003). *Unbewusste Äußerungen und das Verbot des Selbstbelastungszwangs*. Frankfurt/Main: Peter Lang.

Grünwald, G. (1966). Beweisverbote und Verwertungsverbote im Strafverfahren. *Juristenzeitung*, 489-501.

Honts, C. R. (2004). The psychophysiological detection of deception. In P. A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (103-123). Cambridge: Cambridge University Press.

Kassin, S. M. (2004). True or false: "I'd know a false confession if I saw one". In P. A. Granhag & L. A. Strömwall (Eds.), *The detection of deception in forensic contexts* (172-194). Cambridge: Cambridge University Press.

Kassin, S. M., & Sommers, S. R. (1997). Inadmissible testimony, instructions to disregard, and the jury: Substantive versus procedural considerations. *Personality and Social Psychology Bulletin, 21*, 893-898.

Kozel, F. A., Johnson, K.A., Mu, Q., Grenesko, E.L., Laken, S.J., & George, M.S. (2005): Detecting Deception Using Functional Magnetic Resonance Imaging. *Biological psychiatry, 58*, 605-613.

Kozel, F.A., Johnson, K. A., Grenesko, E.L., Laken, S.J., Kose, S., Lu, X., Pollina, D., Ryan, A., & George, M.S. (2009). Functional MRI detection of deception after committing a mock sabotage crime. *Journal of Forensic Sciences, 54 (1),* 220-231.

Kühne, H. H. (2010). *Strafprozessrecht* (8th ed.). Heidelberg: C.F. Müller Verlag.

Langleben, D. D., Loughead, J. W., Bilker, W.B., Ruparel, K., Childress, A.R., Busch, S.I., & Gur, R.C. (2005). Telling truth from lie in individual subjects with fast event-related fMRI. *Human Brain Mapping, 26,* 262-272.

Langleben, D. D., Dattilio, F. M., & Guthei, T. G. (2006). True lies: delusions and lie-detection technology. *The Journal of Psychiatry and the Law, 34*, 351-370.

Moreno, J.A. (2009). The future of neuroimaged lie detection and the law. *Akron Law Review, 42(3),* 717-737.

Moriarty, J. C. (2009). Visions of deception: Neuroimages and the search for truth, *Akron Law Review, 42 (3)*, 739-761.

National Research Council. (2003). *The polygraph and lie detection*. Washington D.C.: The National Academies Press.

Nose, I., Murai, M., & Taira, M. (2009). Disclosing concealed information on the basis of cortical activations. *NeuroImage, 44*, 1380-1386.

Offe, H., & Offe, S. (2004). Experimentelle Untersuchung zur Theorie der Vergleichsfragen in der physiopsychologischen Täterschaftsdiagnostik. *Monatsschrift für Kriminologie und Strafrechtsreform, 87*, 86-102.

Pavlidis, I., Eberhardt, N. L., *&* Levine, J. (2002). Seeing through the face of deception. *Nature, 415, (6867),* 35.

Priori, A., Mameli, F., Cogiamanian, F., Marceglia, S., Tiriticco, M., Mrakic-Sposta S., Ferrucci, R., Zago, S., Polezzi, D., & Sartori, G. (2008). Lie-specific involvement of dorsolateral prefrontal cortex in deception. *Cerebral Cortex*, *18,* 451-455.

Putzke, H., Scheinfeld, J., Klein, G., & Undeutsch, U. (2009). *Zeitschrift für die gesamte Strafrechtswissenschaft, 121*, 607-644.

Rill, H.-G. (2001). *Forensische Psychophysiologie: Ein Beitrag zu den psychologischen und physiologischen Grundlagen neuerer Ansätze der „Lügendetektion“*.
Retrieved from http://deposit.ddb.de/cgi-bin/dokserv?idn=962727717.

Rogall, K. (2010). In J. Wolter (Ed.), *Systematischer Kommentar zur Strafprozessordnung. Mit GVG und*

EMRK (§ 136a). Köln: Carl Heymanns Verlag.

Schauer, F. (2010). Neuroscience, lie-detection, and the law. Contrary to the prevailing view, the suitability of brain-based lie-detection for courtroom or forensic use should be determined according to legal and not scientific standards, *Trends in Cognitive Sciences, 14* (3), 101-103.

Scheffler, U. (2008). In M. Heghmanns & U. Scheffler (Eds.), *Handbuch zum Strafverfahren* (593-874). München: C.H. Beck.

Schmerber vs. California 384 U.S. 757 (1966).

Schmitt Glaeser, W. (2000). Big Brother is watching you – Menschenwürde bei RTL2. *Zeitschrift für Rechtspolitik*, 395-402.

Schneider K. (2010). *Der Einsatz bildgebender Verfahren im Strafprozess*. Lohmar-Köln: Josef Eul Verlag.

Schwabe, J. (1979). Rechtsprobleme des „Lügendetektors“, *Neue Juristische Wochenschrift*, 576-582.

Seiterle, S. (2010). *Hirnbild und Lügendetektion. Zur Zulässigkeit der Glaubwürdigkeitsbegutachtung im Strafverfahren mittels hirnbildgebender Verfahren*. Berlin: Duncker & Humblot.

Smt. Selvi & Ors. vs. State of Karnataka, Criminal Appeal No. 1267 of 2004, Supreme Court of India 05.05.2010. Retrieved from http://www.scribd.com/doc/30969546/Narco-Analysis-Test-Guidelines-by-Supreme-Court-Selvi-v-State-of-Karnataka-2010.

Spence, S. A. (2008). Playing devil's advocate: The case *against* fMRI lie detection. *Legal and Criminological Psychology 13*, 11-25.

Spranger, T. M. (2009). Der Einsatz neurowissenschaftlicher Instrumente im Lichte der Grundrechtsordnung. *JuristenZeitung*, 1033-1040.

Steller, M. (1987). *Psychophysiologische Aussagebeurteilung: Wissenschaftliche Grundlagen und Anwendungsmöglichkeiten der „Lügendetektion“*. Göttingen: Hogrefe.

Stoller, S. E., & Wolpe, P. R. (2007). Emerging neurotechnologies for lie detection and the fifth Amendment. *American Journal of Law & Medicine, 33*, 359-375.

Thompson, S. K. (2007). A brave new world of interrogation jurisprudence? *American Journal of Law & Medicine, 33*, 341-357.

US vs. Semrau, No. 07-10074 MI/P. Retrieved from http://lawyersusaonline.com/wp-files/pdfs-2/us-v-semrau.pdf.

Verrel, T. (2001). *Die Selbstbelastungsfreiheit im Strafverfahren: Ein Beitrag zur Konturierung eines*

überdehnten Verfahrensgrundsatzes. München: C.H. Beck.

Verwaltungsgericht Neustadt (1993), *Neue Zeitschrift für Verwaltungsrecht, 99.*

Vrij, A. (2008). *Detecting lies and deceit. Pitfalls and Opportunities* (2nd ed.). Chichester: Wiley-Interscience.

Watzlawick, P. G., Beavin, J. H., & Jackson, D. D. (1985). *Menschliche Kommunikation: Formen, Störungen, Paradoxien* (7th ed.). Bern: Huber.

Wistrich, A. J., Guthrie, C., & Rachlinski, J. J. (2005). Can judges ignore inadmissable information? The difficulty of deliberately disregarding. *University of Pennsylvania Law Review, 153*, 1251-1345.

Zwiehoff, G. (2000). *Das Recht auf den Sachverständigen*. Baden-Baden: Nomos.

Chapter 4
If man's true palace is his mind, what is its adequate protection? On a right to mental self-determination and limits of interventions into other minds

Jan Christoph Bublitz
University of Hamburg
Faculty of Law
✉ christoph.bublitz@uni-hamburg.de

Abstract While the ethical and legal issues of using neurotechnologies to change one's own mind have been widely discussed, the legitimate ways of changing minds of other persons outside of therapeutic contexts remain vague. Interestingly, the protection of the mind is a rather blind spot in legal thinking. Neuroscience may change this as it provides means to both intervene into minds and measure changes. In this paper, I shall outline why current legal provisions cannot adequately capture interventions into other minds and suggest the recognition of a mind-protecting right, a right of mental self-determination. Furthermore, I propose some elements of a framework of illegitimate interventions into other people's minds based on a normative dualism between direct and indirect interventions, contrasting parity principles between interventions recently put forward in neuroethics.

Keywords neurolaw, brain interventions, freedom of mind, parity principle, normative dualism

Introduction

Increasing knowledge about brain processes and technological advances facilitate new means for interventions into the brain targeting mental phenomena. From deep brain stimulation (DBS), transcranial magnet stimulation (TMS) to pharmaceuticals, tools for changing the inner world of persons are already at hand, and may become more effective and widely available in the near future. These neurotechnologies pose various challenges for society and the law. One of the central issues for neurolaw is to find principles framing the regulation of their use. The ongoing debate over neuroenhancement to improve cognitive capacities relates to one side of the question: What are the legitimate ways of changing *one's own* mind? In this paper, I want to address the other side: What are the legitimate ways of changing the minds of *others*? More precisely, the minds of mentally healthy adults who have neither requested to nor approved of having their minds changed. Therefore, I shall not be concerned with consented interventions (e.g. for therapeutic purposes) but with the broader and very general, yet somehow neglected question over the legal limits of

intervening into other people's minds in the absence of any specific normative relation between interveners and affected persons.[34]

The issues of transforming one's own and changing others' minds share common ground in the question of whether neurotechnological means differ from traditional ways. In an abstract perspective, changing minds is anything but a novel phenomenon. People have always sought to alter their mind states, from ingesting alcohol, chewing coca leaves to education and books. For some, neurotechnologies are the continuation of humankind's quest for optimising cognitive capacities (Galert et al., 2009; Greely et al., 2008), while others point to categorical differences between means (Sandel, 2007), which raises the normative question about their ethical or legal ramifications. But even the first step, describing the differences accurately, is a task harder as it may initially appear. While often taken for granted, Levy (2007) tries to capture them in some detail:

> There are two basic ways to go about changing someone's mind. What we might call the traditional way involves the presentation of evidence and argument…Of course, there is a sense in which presenting evidence is a kind of (indirect) manipulation of the brain – it alters connections between neurons, and might contribute, in a very small way, to changing the morphology of the brain…But direct manipulation of the brain differs from indirect in an extremely significant way: whereas the presentation of evidence and argument manipulates the brain via the rational capacities of the mind, direct manipulation bypasses the agent's rational capacities altogether. It works directly on the neurons or on the larger structures of the brain (Levy, 2007, p. 69).

Having said that, Levy goes on to argue for an ethical parity principle: *"our new ways of altering the mind…ought not to be regarded, as a class, as qualitatively different in kind from the old"* (Levy, 2007, p. xii) In its weak version, the parity principle claims *"unless we can identify ethically relevant differences between internal and external interventions and alterations, we ought to treat them on a par… [T]he mere fact that one kind of intervention is internal is not a ground for objection"* (Levy, 2007, p. 62). Levy does not claim we should not worry about new means of mind interventions but that we should not be especially worried about *"internal means of manipulating the mind, not until far more powerful techniques come into existence"* (Levy, 2007, p. 144) since traditional means might be equally effective: *"[T]he kinds of powers that neuroscience promises in the near future pale in comparison to the mind…control techniques already in existence, in power and in precision"* (Levy, 2007, p. 144).

Levy's remarks are very helpful for distinguishing interventions. Regarding the normative claim,

[34] Evaluating the legitimacy of interventions in special relations such as doctor-patient or state-citizen requires a model of legitimacy in standard cases. Furthermore, between changing one's own or another person's mind lies the third category of changing another's mind by request in which different normative criteria obtain (Merkel, 2007).

however, I shall argue to the contrary. Concerning interventions into other minds, a legal parity principle between interventions cannot obtain, not even prima facie. Rather, I propose a *normative dualism of interventions*: Without consent, the law should prima facie consider indirect interventions permissible, direct interventions prohibited. Nonetheless, I concur with Levy that worries over new technologies may prompt us to reconsider some traditional means of changing minds too. Let us take a look at some ways we change other persons' minds, how the law deals with these interventions de lege lata (in its current state) and why it has problems to adequately capture mind interventions. As a solution, I propose to recognise a right to mental self-determination de lege ferenda, outline its contours and a framework of illegitimate mind interventions.

Changing the minds of others

As said, it is a truism that we change each other's minds all the time. Every act of communication changes the minds of speaker and listener, and, as we may reasonably assume, changes brain processes as well. While I am writing these words, I expect to change the mind and the brain of you, the reader. I even intentionally strive to change it. Obviously, these kinds of intervention into another's mind are beyond any ethical or legal concern.

In contrast, other interventions are clearly impermissible. Last year, scientists voiced worries over security weakness of neuro-devices like DBS, which may enable hackers to change their functioning (Kohno, Denning, & Matsuoka, 2007). Is there anything closer to mind-control than remotely controlled DBS-'mind hacks', in quite a literal way? But even without hacking, taking over control of neuro-devices by force in traditional ways allows interveners to change minds of others quite invasively. Imagine scientists turn a patient's DBS on and off modulating his moods from moment to moment as they please to observe his reactions for a scientific experiment. Surely, without consent this is illegitimate, and, by the way, not because of the illicit use of machine and property, but because of the mental effects. Likewise, changing moral reasoning through TMS as recently demonstrated in a study (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010) appears illegitimate.

Other interventions are harder to assess. Consider oxytocin, a neuropeptide recently gaining popularity in the press and neuroethical thought experiments. Its name derives from Greek (roughly: sudden birth) since it is released during labour. Oxytocin seems to impact various interesting mental and behavioural properties from bonding, maternal behaviour and sympathy to trust and risk-taking (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005; Fehr, Baumgartner, Heinrichs, Vonlanthen, & Fischbacher, 2008). Some have exaggeratedly called it the "molecule of love". So, if I, seeking to win your sympathy, spray some odourless oxytocin in the room before a talk, or a business partner before negotiating a risky deal – is it illegal? Do I commit a crime? Parts of the answer surely rely on its efficacy. However, I will set aside empirical aspects and concentrate on normative issues assuming interventions to contribute to mind-change.

Also, take an example Dennett put forward: *"Consider the case of the eloquent philosopher who indirectly manipulates a person's brain by bombarding his ears with words of ravishing clarity and a host of persuasively presented reasons, thereby inducing all his desires, beliefs, and decisions"* (Dennett, 1984, p. 64). Intending to ridicule frequent worries over manipulation and free will (an issue I will touch upon) Dennett

(1984) asks whether there is anything dubious in the philosopher's way of changing minds.

And finally, consider a phenomenon more common than desire-inducing philosophers: marketing. Boasting claims over the efficiency of marketing methods developed with the help of functional Magnetic Resonance Imaging (fMRI) data can be found throughout the rapidly expanding neuromarketing literature. Although these claims tend to be examples of good self-marketing rather than good science and their explanations, e.g. the reverse inference from brain states to mental states, are often overly simplistic (Ariely & Berns, 2010; Pauen, 2007), it is important to keep in mind that the neuromaketer's explanandum is not so much decision-making but ways of influencing it. As medicine aptly proves, changing phenomena does not require detailed understanding of its underlying mechanisms. On a larger scale, psychological surveys indicate that marketing works, but no one knows exactly how and why. In light of new findings over its efficacy and involved mechanisms the limits of legitimately changing others' preferences through marketing may need to be reconsidered.

Surveying the normative landscape

What does the law de lege lata have to say about the legitimacy of these interventions, from DBS to neuroscientifically optimised ads (hereforth: example interventions)? Let us take a closer look at the existing landscape of norms relating to mind interventions. I hasten to add that there is, of course, not *the* law, but rather various different legal systems on different levels, from civil to criminal, from local to international law (the latter I call jurisdictions). Norms and doctrines may vary from one jurisdiction to the next, sometimes legal cultures are more divergent than ethical schools of thought. Even a systematic analysis of all norms relating to the mind in one jurisdiction is a *Herculean* task no one has, as far as I see, ever pursued. Therefore, the following remarks have to remain on a rather abstract level and resort to some basic principles of the law without commitment to a particular system. Apparently, cognitive liberty and its relation to the 1st amendment have been debated for the US legal system (Boire, 2003) with which I am not familiar enough to comment on. However, even there no legal theory of protection of the mind seems to exist. Thus, I hope my claims might apply to many jurisdictions.

In their abstract formulations, legal provisions regularly do not refer to particular types of actions, e.g. spraying oxytocin, but to certain states of affairs (the protection of bodily integrity prohibits any action which causes bodily harm). Therefore, it is helpful to distinguish mind interventions by their outcomes. Two types appear morally and legally problematic: interventions causing psychological harm and those altering will-formation and preference structures.

Protection of bodily integrity

Every jurisdiction provides some kind of protection to bodily integrity; inflicting bodily harm constitutes a criminal offence throughout the world. Do the example interventions cause bodily harm? Being exposed to an electromagnetic field via TMS surely produces changes in the brain, pharmaceuticals modify levels of neurotransmitters and watching ads changes brain activity. But do these changes constitute *harm* to the body? Criminal provisions often define infliction of bodily harm as actions having a detrimental effect on

bodily tissue, substance or the functioning of organs. Mere changes of the particles comprising the body are insufficient, since, on a closer look, at the microscopic level bodily changes occur all the time, yet it is hard to conceive of the body as being in a permanent state of injury.

Thus, the negative effects on the material side of the body have to exceed a threshold. Some of the old forms of mind interventions like psychosurgery or electroconvulsive therapy certainly constitute bodily harm due to their negative bodily effects. Therefore, without prior informed consent, they violate the right to bodily integrity. With TMS and DBS (once the device has been surgically implanted) it depends on facts of the concrete case, especially what kind of, if any, damage has been caused to nearby brain tissue. Psychopharmaceuticals may constitute bodily harm if they have negative bodily side-effects. Their main mind-altering feature, however, the change of levels of neurotransmitters in synaptic clefts, does not constitute bodily harm. If it was, at least within a normal range, every act of communication with similar effects may constitute infliction of bodily harm - an absurd consequence. Hence, mind interventions cannot be adequately captured on the bodily level. The law has to recognise the mental side, and in fact, some jurisdictions have considered extending the protection of bodily integrity to mental phenomena.

German courts hold that mental harm can constitute a violation of bodily integrity if they *"manifest somatically"*. This may sound progressive, but an analysis of judgments reveals that the protection of the mind is restricted to cases involving harm to the body (Bublitz, 2011). Furthermore, one should be aware of the fact that legal reasoning is deeply permeated with mind-brain dualisms, to be witnessed in a widespread distinction between physical and emotional harm. 'Pure' psychological harms, those not accompanied by damage to the body's tissue, are not considered as bodily harms. Obviously, such interpretations rely on mind-brain dualisms of a stronger sort presupposing the existence of purely psychological harms and purely mental states. In light of current debates in philosophy of mind such a position is hard to maintain. More than likely, all mental states are caused, realised by or at least supervene on bodily states.

In the UK, the House of Lords held:

> The phrase 'actual bodily harm' is capable of including psychiatric injury. But, it does not include mere emotions such as fear or distress or panic nor does it include, as such, states of mind that are not in themselves evidence of some identifiable clinical condition.[35]

Thus, the protection of bodily integrity covers, at best, the infliction of mental disorders. None of our example interventions would meet this threshold.

In tort law**,** the psyche is afforded broader protection. In their seminal paper on privacy the US scholars Warren and Brandeis (1890) noted:

[35] Reg. v. Chan-Fook [1994] 1 W.L.R. 689; Reg.v.Ireland/Burstow [1998] AC 147.

> [O]ur law recognises no principle upon which compensation can be granted for mere injury to the feelings. However painful the mental effects upon another of an act, though purely wanton or eyen malicious, yet if the' act itself is otherwise lawful, the suffering inflicted is *damnum absque injuria* [a loss without injury, a damage without a violation of someone's rights]. Injury of feelings may indeed be taken account of in ascertaining the amount of damages when attending what is recognised as a legal injury, but our system, unlike the Roman law, does not afford a remedy even for mental suffering (Warren & Brandeis, 1890, p. 197).

Today, I suppose, most jurisdictions do recognise tort claims over psychiatric injury to some extent. However, requirements are quite high. In the US, claims for *"intentional infliction of emotional distress"* require conduct to be outrageous and mental distress incurred to be severe. Several European jurisdictions stipulate mental injuries have to be caused by *"nervous"* or *"sudden shocks"*. Therewith, courts try to curb the virtually limitless claims over psychological harms (McInerney, 2009). Also, harms have to be quite severe, so severe, in fact, that posttraumatic stress disorder is contested in various jurisdictions. It is one of the hopes expressed by some legal scholars that neuroimaging may be employed to estimate the amount of pain or suffering inflicted in order to grant appropriate compensation. For epistemological reasons, I suspect this hope is farfetched since it tends to ignore the central point in the reductionism debate. If qualia, the subjective experience, the 'what-it's-likeness', cannot be reduced to brain states, then the intensity of discomfort cannot be read off their images.[36] But this leads us astray.

We can observe that courts only reluctantly grant remedies for mental injuries without somatic harms, but on the bodily level alone, negative mental changes are hard to be identified. Evaluating mind interventions reasonably requires taking the mind seriously – which implies recognising the mind in 'its own right' and discussing to which mental phenomena protection should be afforded. Before the procedural aspects of neuroimaging in the courtroom can be assessed, the role of mental states in substantive law has to be reconsidered.

Privacy/personality rights

A different set of rights found in most constitutions and human rights treaties protects privacy or personality.[37] The central idea of privacy rights is that individuals are entitled to private spaces free from

[36] At least, when it comes to inter-subjective comparisons between pain-intensity, epistemological obstacles seem irresolvable, Kolber (2007) is more optimistic.

[37] Note that rights enshrined in constitutions or international human right treaties are primarily designed to apply in the vertical relationship between state and citizen, not horizontally, in private-private relations. They give rights against the state, e.g. to protect and not to violate privacy, but not against other citizens. For horizontal, third-party applicability they have to be transformed by further regulations, e.g. by recognising a tort of privacy. I will leave out such details and

unwanted intrusion. Traditionally, the protected sphere is the home in contrast to public places. In a classic understanding, privacy provisions only guarantee a right to be let alone, protecting a *Thoreauean* like seclusion in solitude. From the prohibition of surveillance measures based on privacy analogies can be drawn to mental privacy, barring neuroimaging or mindreading. However, in this sense privacy seems to lack the resources to deal with mind interventions in public social interactions. Yet this is what we may be most interested in. But as the notion of privacy has been steadily extended covering areas as anonymity, reputation, false press statements and Internet data, this line can be further expanded to the mind. After all, in the words attributed to the artist *Joseph Beuys*: The true palace of man is his mind.

Personality rights protect the integrity and expression of one's personality, classic examples include the right to sexual identity or self-definition of personality elements disclosed to the public. The scope of personality rights often overlaps with privacy rights. Interestingly, the notion of personality seems to remain largely unexplored in legal thinking. Although the German Constitutional Court held that Art. 2 of the German Basic Law confers absolute protection to the *"core elements of the personality"*, which means that interferences may not be justified by appeal to opposing rights, it has never defined what a personality or its core elements consist of. In a narrow understanding, personality traits are somewhat constant and consistent properties closely connected to that 'what binds the person's inner core together'. So it remains unclear whether mind interventions altering preferences, behavioural dispositions or emotional propensities of rather trivial sorts qualify as changes in the personality. Presumably, a broader meaning of personality needs to be construed to capture transient changes in mood or preferences caused by oxytocin, TMS or reading these words.

Art. 8 of the European Convention on Human Rights (EC) guarantees a *"right to respect for private and family life…"* which the European Court of Human Rights (ECHR) has interpreted quite extensively covering not only privacy but also the development of one's personality in *interaction* with others and conferring a right to *"psychological integrity"* which has recently joined the rank of fundamental rights in the EU.

Mental integrity

Almost ten years after its proclamation, the Treaty of Lisboa set in effect the Charter of Fundamental Rights of the European Union (EUCh) in December 2009. The Charter is the EU's first codified 'bill of rights'[38] Art. 3.1 reads: *"Everyone has the right to respect for his or her physical and mental integrity"*. Art. 3.1

---

assume horizontal applicability.

[38] This may sound confusing to non-lawyers (and non-Europeans): The European Convention on Human Rights is an ordinary international treaty adjudicated by the European Court of Human Rights in Strasbourg, France. Violating the

is one of the few norms explicitly protecting the mind. As it is neither contained in the EC nor in most national constitutions, the term caused some controversy during final debates. To various delegates *"mental integrity"* appeared unfamiliar and its scope unclear. A Swedish representative declared: *"The concept of mental integrity remains a mystery to Swedish experts who cannot find it in either the Convention on Biomedicine and Human Rights or any other international instrument"* (Borowsky, 2006, p. 120). The term was adopted, nevertheless, since Art. 3 was designed to afford protection against all conceivable biotechnological interventions.

Art. 3 does not indicate what mental integrity means and leaves its interpretation to the courts. For some commentators it is synonymous with mental health. Others suggest that Art. 3 should also afford protection against coerced psychiatric treatments which may improve health but interfere with integrity. Even others consider it more broadly conferring protection against indoctrination or brainwashing (Rengeling & Szczekalla, 2004). Conceptually, it is hard to conceive what the *"integrity"* of a particular mental state is supposed to mean. Either a person is in or has a particular mind state – e.g. a thought or an emotion – or not. The same is true for 'the mind' – what is a disintegrated mind? Some forms of mental illness might be called 'disintegrated mind states', but the purpose of Art. 3 is to cover more than those exceptional states. Alternatively, it could protect the mind in its current state, the mental status quo. But this is quite unconvincing in light of the fact that the mind is in an ever-changing 'stream of consciousness'. Thus, the concept of *"integrity"* cannot be easily transferred from the rather static body to the dynamic mind. Mental integrity is a metaphor in need of further explication. Whether any of our example interventions interfere with it remains an open question. Ambiguities notwithstanding, the drafters of the Charter send a clear message for stronger protection of the mind, which will reverberate in the member states' legal systems.

Freedom of thought/freedom of speech

Another important provision for mind interventions is the right to freedom of thought, enshrined in every human right treaty (Art. 18 Universal Declaration of Human Rights – UDHR, Art. 18 International Covenant on Civil and Political Rights – ICCRP, Art. 9 EC, Art. 10 EUCh). Freedom of thought is one of the strongest, most fundamental rights, demonstrated by the fact that it is protected unconditionally. Neither the

---

Convention is a violation of an international treaty. The ECHR has no direct powers to enforce compliance with the Convention, however, parties to the treaty usually incorporate rulings into domestic law. The European Union is, arguably, a state on its own to which the member states have conceded some of their sovereign powers. The EU has a parliament, executive and judicative branches, but limited areas of competency. The highest Court of the EU is the European Court of Justice (ECJ). The Charter of Fundamental Freedoms only applies to acts of the EU, not of member states. Nevertheless, due to the process of legal harmonisation throughout Europe, the Charter will influence court rulings in member states.

Universal Declaration of Human Rights, nor the EC provide limitation clauses. Art. 18 ICCPR & UDHR do *"not permit any violation whatsoever on the freedom of thought"* (UN General Comment No. 22, 1993).

Freedom of thought is closely related to freedom of speech and expression and in philosophy, literature and even the law, they are often used interchangeably. For legal purposes they should be kept apart. Freedom of thought refers to persons' inner sphere (called 'forum internum') while freedom of speech and expression entail actions in the outside world, the manifestation of beliefs (forum externum). The forum internum also covers religious beliefs and personal moral conscience and is often regarded as inviolable. So even if thought was nothing more than inner speech it is protected unconditionally, only its external expressions are restrictable. Sometimes, the right to freedom of speech is deemed to include processes prior to speaking, e.g. forming and having opinions. In this spirit, Art. 19 UDHR & CCPR guarantee a right *"to hold opinions without interference"* which overlaps with freedom of thought.

So, high-ranking rights cover mental phenomena as thought or belief-formation. In sharp contrast to their omnipresence, visibility and declared indispensability for democratic societies stands their lack of practical importance. For instance, in his over 50 years of existence the ECHR has only decided a handful of cases regarding freedom of thought, none of them bearing any relation to our present concern. Compared with the freedoms of religion and conscience protected by the same article, freedom of thought is an almost empty declaration.[39] There are no definitions over its meaning, scope or possible violations. Even commentators rarely attempt to define what this fundamental liberty might encompass. Some provide examples of violations: again, brainwashing or indoctrination (Vermeulen, 2006). To others it prohibits *"any conduct intended to change a thinking process, change an opinion or force divulgence of conviction"* (Clayton & Thomlinson 2000, p. 976). Plainly, this must be wrong – otherwise, every university has to be shut down immediately. Given this ambiguity, one wonders why the ECHR never took a chance to formulate an idea of what freedom of thought may mean. Perhaps there was no necessity as all cases presented to the court were resolvable by other articles. However, as other articles, are limited (Art. 10 II ECHR), the logic of the law would demand testing complaints against the strongest right possible.

Thus, we can only speculate over the scope of freedom of thought. Narrowly construed it might encompass only the capacity to think, i.e. the cognitive processes bringing about conscious mental phenomena which classify as thoughts. But does it cover the act of thinking or merely having a thought? And what is a thought, a subclass of conscious mental phenomena? Do thoughts include phantasies and emotions, do they need semantic content? And what is the act of thinking – consciously processing information? What about being in pain, loving, associating pictures with music or dreaming? And, after all, how can mental phenomena of whatever kind be unfree? Have we always thought freely?

[39] Interestingly, the constitutions of EU member states do not protect freedom of thought, hence, there are no national court cases from which interpretations could be drawn (Bernsdorff, 2006).

I suspect legal scholars and courts are not too ambitious to get into these questions and rather cling to the belief that freedom of thought is not only legally, but factually inviolable as thoughts and the mind are 'intangible' and beyond the reach of interventions. During the drafting of the German Constitution after the 2nd World War, delegates remarked:

> Thoughts, thank god, are shielded by the corpse and can neither be accessed nor restricted by anyone. It is unnecessary to protect freedom of belief – who could ever touch upon them? A norm: 'thoughts are free' is meaningless, its content self-evident, since no one could ever interfere with them. Therefore, such a norm would be without application and merely an empty phrase (Kahl & Thoma as cited in Faber, 1977, p. 48).

In some opposition to the presumption of the inviolability of thought stands the fact that practices as brainwashing – whatever it is – or psychoactive substances altering contents of thoughts, generating associative thinking, accelerating or slowing down thinking or modifying patterns of thought, etc. have a long history. Likewise, psychological findings have for at least a century indicated that mental phenomena can be changed in various and dubious ways. Somehow the law has not caught up with these insights and remains committed to a rather naïve, first-person perspective impression that thought is as a matter of fact free. Do our example interventions violate freedom of thought? By current (ill-defined) standards I suppose not as long as affected persons maintain capacities for thinking, making up their minds and weighing arguments.

Other provisions

Finally, it should be noted that some types of mind interventions fall under special regulations, e.g. oxytocin is included in drug-regulations prohibiting their unlicensed distribution and application. Presumably, among the many goals these regulations pursue is a rather broad understanding of the potential harmful effects of these substances. So, in a weak sense these regulations afford protection of the mind, even though they are not founded on a principled recognition of mental self-determination (at times, drug regulations even interfere with it). Furthermore, severe mind interventions can amount to torture or inhuman and degrading treatment (Art. 3 EC). Arousing *"feelings of fear, anguish, inferiority and humiliation"* can violate Art. 3 if mental suffering is intense (*Keenan v. UK*, ECHR 272229/95). Famously, the ECHR held that extraditing criminals facing the death penalty to the US may amount to torture – not because of the punishment, but because of the time waiting in death row (*Soering v. UK*, ECHR 14038/88; Lillich, 1991). While there has been some success in the prevention of physical ill-treatment, the non-physical forms of torture have been somewhat neglected (Neziroglu, 2007). It remains unclear which methods aiming at mental distress constitute torture as recently and infamously demonstrated by the war-on-terror practices endorsed by the former US president. Without any doubts, mock executions are torture. However, psychological torture is not our concern here.

Preliminary conclusion

In criminal law, the protection of the person is focused on injury to the body. From unlawful touching to

medical treatments contra legem artis, jurisdictions have developed detailed rules over (im)permissible conduct with other persons' bodies. Concerning the mind, things are different. With the exception of severe mental injuries, the law remains largely silent about normative principles on altering another person's state of mind. To put it bluntly: Legally, you can mess up another person's mind quite intensively, but you are not allowed to touch her body inappropriately.

Although freedom of thought and opinion or personality rights are central constitutional guarantees, their scope and likewise their violations remain unclear. The usually adduced examples such as brainwashing or mind-control are themselves vague, lacking definitions or case applicability and presumably refer only to extraordinary severe interventions. But are any less invasive interventions per se legitimate? Legal provisions and their interpretations oscillate between being too narrow (mind-control) or too broad (changing opinions). What is missing is a coherent theory of illegitimate mind interventions providing guidelines for concrete cases.

I have elaborated on rights and their interpretation so extensively in order for lawyers to recognise that, contrary to widespread belief, there are gaps in current doctrines and for ethicists to see how the law argues hoping to connect disciplines. Albeit intuitively it may seem feasible to discern morally permissible and impermissible mind interventions without a general theory of mind protection, the law deals with rights, their scopes and limits. Neurolaw is not just an (codified) extension of neuroethics. Norms pertaining to the mind have to be coherently aligned with the overall system of rights and duties, and legal theory should attempt to define principled criteria under which cases can be subsumed. Simply outlawing manipulations or causing mental harm stipulating a neminem leadere principle for the mind is not a feasible option given the fact that persons seek to influence, hurt and cause each other sadness and sorrow all the time. Unless one seeks to turn society into a community of superficial, false and faked kindness, the law has to acknowledge that social life consists of clashing opinions and at times it hurts.

There are manifold additional reasons why the mind is a rather blind spot in the law. Historically, classic concepts of law such as Kant's (1797) restrict its purview to the regulation of behaviour in the external world because only there spheres of freedom can collide (and acting from good will is only a moral, not a legal obligation). Practically, the law has problems with proving mental states. Philosophically, the law and courts rely on some intuitive mind-brain dualism and further, applying notions as causation to the mind is problematic. Politically, one could indeed ask whether the law is the proper tool for regulating inevitable conflicts of interpersonal relationships. Presumably, there is more than a grain of truth in all these observations. Yet, in light of the example interventions it is anything but self-evident why the law should exercise a 'judicial restraint' on almost all 'matters of the mind'. After all, the law's mission is to protect interests of individuals, and there certainly is an interest not to be exposed to stimuli even when they fall short of brainwashing. Therefore, the challenge for the law in the age of neuroscience is to formulate a modern notion of free thought or freedom of mind affording reasonable legal protection to neuronal and psychological processes underlying and shaping thinking and feeling. Informed by neuroscience, psychology and philosophy of mind, the law has to formulate a framework into which empirical findings can be incorporated. Nonetheless, this framework is essentially normative and has to be derived from legal considerations of which I will outline some in the following.

A right to mental self-determination

When the natural boundaries of the mind, the skull, can be surmounted by neurotechnologies, normative boundaries have to be established. Thus, the law should recognise a mind-protecting right. In codified-law jurisdictions this means ultimately drafting and passing a bill; alternatively, existing statutes might be modified within the permissible margins of interpretation. But what is its scope? To begin, it may be useful to detach one's perspective from given categories, notions of mind-control and the like and ponder over the question which mind interventions appear illegitimate. For the law, this is already a paradigm shift, and it may reveal that large shares of mental occurrences are responses to the world enabling interaction with the (social) environment. Unless one strives to attain nirvana and Buddhist detachment from the world, the mind and mental changes are conditions sine-quibus-non for living, and not surprisingly in many jurisdictions the criterion for death is the irreversible ending of brain and mental activity. But what should the law protect given the mind's peculiarities? Mental autonomy, it seems, is a rather misleading idea – we do not give ourselves the laws on which our minds work. Also, protecting mental integrity seems dubious if it implies comparing one mental state to another and finding out whether it worsened as the mind is a highly dynamic system. I think the best abstract formulation of the protected interest is mental sovereignty or mental self-determination, to be understood in opposition to heteronomous influences on mental functions. Let us take a closer look what this may mean in a legal context.

The scope of the right

The right to mental self-determination (SD) should encompass all mental states, i.e. not only thoughts or opinions, but also emotions, behavioural dispositions, even subconscious processes. Drawing meaningful distinctions between types of mental states is impossible, even the term mental 'state' cannot be taken literally. As said, the mind is an ever-changing dynamic system, and presumably all mental states are functionally interdependent. Here, some words about emotions are in order. I presume (negative) feelings are actually part of the mind's working and serve functions. They alert us and guide behaviour, according to Damasio's (1994) somatic marker hypothesis, they are even necessary for rational decision-making and hence inducing negative feelings per se is not a sign of a violation but of a well-working system. Only in extraordinary cases inducing such feelings can be considered a 'damage' or harm in a legal sense. Particularly, a distinction has to be drawn between appropriate and somehow inappropriate emotional responses to the world elicited by mind interventions (e.g. in the DBS case). However, since humans are particularly weak in controlling emotions, sometimes controlling other people's emotions or mood via evoking appropriate responses may be the most efficient way to control their thinking and decision-making.

Furthermore, as we have seen, the law protects the personality of individuals. But what is a personality? I suspect it can be broken down into a web of desires, beliefs, emotional propensities, etc. and hence, a personality is a conglomerate of interwoven mental states seen from an abstract perspective. Therefore, it should be possible to combine the protection of thoughts, opinions and personality into one unified right, rendering further distinctions obsolete. Prima vista, all mental states should be protected equally.

The right to mental SD has several dimensions: It is a liberty right, permitting right-holders (every person) to freely exercise their mental capacities and change their mental states placing obligations onto others not to interfere therewith. Regarding neuroenhancement, the legal question arises whether the right also entitles to transform mental states with the help of neurotechnologies. Prima facie, I suggest it does – however, employing neurotechnologies may be restricted to a larger extent than exerting mental capacities or mental actions. The latter, I claim, cannot be legally restricted at all as long as they exclusively concern the forum internum. For present purposes, however, we are only interested in the negative dimension of the right, i.e. which kinds of obligations it places on others. Which conduct interferes with and possibly violates mental SD?[40] Put simply, interventions severely undermining mental self-determination are illegitimate. Defining which interventions severely undermine mental SD requires a firmer understanding of what it is – or at least, what it is not. Self-determination is an ambiguous notion and may raise some objections.

At the outset, an objection which is metaphysical in nature and supposedly supported by findings in neuroscience and psychology should be addressed: There is no 'self' and therefore no self-determination, so the question whether it has been undermined is meaningless. Speaking of self-determination implies dubious metaphysical assumptions of an 'I', a 'self' or controlling-entities which are based on the homunculus fallacy, a *"fatal theoretical error"* (Wegner, 2005, p. 19).

Before the law joins in discarding notions as self-determination altogether, it is important to see in which ways they might be fallacious and in which not. Perhaps there is no transcendental ego, no immaterial 'I' hovering above the empirical world. Yet, for present purposes, we do not need to rely on homunculi. Mental self-determination is not a descriptive, but an ascriptive notion. It is a normative judgment based on observations from both first- and third-person perspectives relating not so much to freedom from natural or neuronal processes, but from interference by others.

Without appealing to dubious entities, it is possible to ascribe to individuals mental capacities such as remembering, concentrating, calculating, language skills, logical reasoning, shifting attention or pulling oneself together, etc. Notwithstanding contested issues as mental causation, persons can be described as having some sort of conscious control over their mental life. Interventions can undermine these control capacities and thereby interfere with the right to mental SD. This is the more active component of self-determination.

Nonetheless, it is easy to forget how limited capacities for self-control are. It is a striking fact that not even our beliefs are under the free disposition of our wills (Pettit & Smith, 1996; Noordhof, 2003). For instance, I always found reincarnation a comforting vision, yet, as much as I want, I simply cannot bring myself to believe in it – try for yourself. From what enters our stream of consciousness or how we feel, which

[40] A terminological note: I speak of interference when actions fall within the ambit of mental SD. Therefore, they are prima facie illegitimate and in need of justification. If they cannot be justified, they violate mental SD.

moments we remember and what we forget to limits of introspection and motivation, we are to a large degree more passive bystanders to inextricable psychological forces. Even trying not to think about something and halting the wandering of the mind takes years of practice. It is a fact of the mental conditio humana that conscious mental powers are quite limited.

What does this mean for a right to mental SD? The law can only protect capacities of self-control to the extent they really exist. If there is no control over certain mental elements, it cannot blame others with having undermined it. Take as a vivid example sexual orientation. If neuroscientists identify its neuronal correlates and means of modification (surely, this is a thought experiment) and change someone's sexual orientation, it is hard to speak of them having undermined control – there was no control to begin with. Nonetheless, their intervention appears impermissible. The reason for this is that self-determination is to be understood broader than conscious self-control. Consciously uncontrollable mental elements still 'belong' to the individual; in fact, they constitute large part of his character. Normatively, mental sovereignty implies that others are not allowed to interfere with these states – this follows straightforward from the postulate of personality protection. The real problem is that in light of such a broad understanding of self-determination, it may seem as if any intervention leaves self-determination intact. Unless persons are 'brains in a vat', hooked up to supercomputers like in famous thought experiments, persons bring about, in a sense, all their mental states themselves. Again, the criterion seems to be either all-encompassing or too-exclusive. Here is the place to draw normative distinctions.

Take the famous example of a person P who decides to raise her arm and then raises her arm.[41] P may feel fully self-controlled. Nevertheless, if we know that X has sent magnetic stimuli to her brain producing both decision and arm movement, it is fair to contend P's self-determination has been undermined, although, strictly speaking, her very own neuronal processes have been involved. A normative judgment allows ascribing the decision to the stimulus. Note how the meaning of self-determination differs here from other current debates, especially in the free will context. There, self-determination is questioned when mental states are determined by neuronal processes. Here, we are interested in a right to mental SD, and rights concern interpersonal relations: Does X respect P's self-determination? I guess not. So, roughly mental SD protects against interventions that undermine control capacities or change elements of the personality.

Mental self-determination and free will

Recognising a right to mental self-determination poses some interesting theoretical challenges for the

[41] The classic experiment by Brasil-Neto and colleagues (1992) concerned fingers. Schleim (2010) reports he could not find any scientific account of induced arm-raising cases in which subjects experienced the decisions as willed by their own.

law. The prime objection against mental SD I often encounter when speaking to legal scholars can be formulated like this: Persons have free will, or at least, the law presumes so. As long as the law maintains this premise it cannot declare an intervention as having undermined self-determination. Inner coherence of the law demands to consider agents as either free (and responsible) or unfree (not responsible). Declaring simultaneously that someone acts freely yet his mental sphere has been interfered with illegitimately appears to be an utter contradiction. Beyond the rare accepted cases in which the law considers agents unfree, the premise of free will leaves (almost) no room for undermined self-determination.

Indeed, mental self-determination stands in an intricate relationship with the free will premise. Due to space constraints, here is my argument in a nutshell: The premise of free will is not inconsistent with a right to mental SD, but on the contrary, presupposes it. A right to mental SD is, in a sense a right to free will. Roughly, the law places on individuals an obligation to arrange their mental sphere in a way that only law-complying conduct arises. Only in exceptional cases the law is willing to acknowledge that persons could – in whatever sense – not have acted otherwise. For this, the law makes numerous metaphysical and empirical assumptions, which have been assailed from many directions for ages. Nevertheless, if the law maintains the free will premise claiming you could have acted, and for that matter wanted otherwise, and leaves to the individuals the internal configuration of their minds but expects them to do so in a certain way and treats them as capable of doing so, then it has to confer to them the legal powers to achieve this, i.e. a right to keep others from interfering with mental SD. Considering persons as self-determined entails granting them the legal powers of self-determination. Therefore, a right to mental SD is a prerequisite, a conditio-sine-qua-non for the free will premise.

From this, we can even infer a first criterion for illegitimate interventions into the minds of others: Mind interventions rendering agents non-responsible are illegitimate. Why? Respecting other persons mental SD implies respecting the other as a subject in the legal sense. Being a legal subject means to be a free, fully right-bearing and responsible agent. As soon as agents are not considered autonomous and not held responsible for their deeds any longer, they lose their status as legal subjects (in this strict sense). The law does not need to recognise their actions as free anymore and if persons remain in a state of legal incapacity for a longer time, the law may even have to assign legal guardians or representatives taking care and making legally binding decisions on their behalf. Often overlooked, responsibility has two inseparably connected sides. The first side is sometimes portrayed in a negative light, i.e. being an apt target for negative sanctions as punishment. It correlates with the other side, i.e. being considered and respected as an autonomous person, which means being allowed to act without legal constraints or interferences.

Why should this status be respected by others? Because, I suppose, it is the basic assumption of any legal community of free and equal persons. In the hypothetical moment in which persons left the stature of nature to enter into a state of rule of law, one of the founding elements of their social contract must have been to accept everyone else as an equal member of the community. This is inherent to the idea of a social contract – actually, it is its central idea, reappearing in Rawls' (1971) two principles of justice: *"Each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others"* (Rawls, 1971, p. 60). Respecting each other as entitled to equal rights means refraining from expelling each other from this status as this is a denial of their right to have equal rights, freedoms and responsibilities. One

cannot simply kick others out of the game without violating the rule establishing it: recognising each other as partners of equal rights. Therefore, as a minimum, we owe each other respect for the mental capacities required for responsibility. Notwithstanding objections levelled against social contract theories, I think this is a fairly sound starting point in a world of moral uncertainty.

Manipulation cases in the free will debate

Speaking of free will and mind interventions, it is inevitable to take notice of a close-by philosophical debate revolving around manipulation cases, which play a pivotal role in the contemporary free will debate. As there is something to be learned from it, here is an admittedly rough excursion into the heart of the controversy. This is a manipulation case borrowed (and slightly modified) from Pereboom (2003):

> Agent A was created by nefarious neuroscientists, who can manipulate him directly through the use of neurotechnologies, but he is as much like an ordinary human being as is possible, given this history. Suppose these neuroscientists manipulate his process of reasoning by which his desires are brought about and modified. The neuroscientists manipulate him by pushing a series of buttons just before he begins to reason about his situation, thereby causing his reasoning process to be rationally egoistic. Due to this, he develops the desire to kill V, which he does (Pereboom, 2003, pp. 112).

The question is whether A is responsible for killing V and the intuitive answer is that he is not. There is something unfair about holding A responsible although he acts in accordance with his desires which he might modify in light of convincing counter arguments (he may even, in Frankfurt's words, wholeheartedly affirm them) and there are no other internal conflicts. Traditional conceptions of (non)responsibility do not capture the problems posed by manipulation cases since agents set ends for themselves and are neither controlled nor constrained or coerced. And here, the debate starts. Further cases are introduced in which the power of the manipulator gradually diminishes. In the end, the neuroscientists vanish and leave behind a world governed by (deterministic) natural laws and persons shaped by social and cultural forces ultimately beyond their control. The challenge is to pinpoint at which stage the judgment shifts and A is considered fully responsible. Some claim there is no relevant difference between cases and thus if A is (not) responsible in the first, he is (not) in the last while others try to make out relevant differences. To cut a long story short, both compatibilists and libertarians (the positions contending that humans sometimes act freely) have difficulties in accounting for the relevant differences as long as they only regard the internal structure of the agent's (A) psyche.[42] Whatever a theory's sufficient conditions for responsibility, one can always come up with a scenario in which they have been installed by nefarious neuroscientists. Therefore several authors introduce

[42] Originally, manipulation cases were designed as an attack on compatibilism (Kane, 1998). However, it seems that libertarian views are susceptible to the same arguments, e.g. Mele (2006).

further conditions of responsibility which pertain to the history of an agent's preferences. Agents have to act on their 'own' or 'authentic' preferences. Unfortunately, most authors do not flesh out this intuitive notion in more detail. Mele (1995) is a noteworthy exception. To him:

> …there is also a negative historical constraint on autonomy which I have called authenticity...A necessary condition of an agent authentically possessing a pro-attitude P...is that it be false that having P...is, as I will say, compelled* – where compulsion* is compulsion not arranged by S...[Sometimes] agents come to possess pro-attitudes in ways that *bypass* their control capacities over their mental lives...Bypassing is sufficient for compulsion...provided that the bypassing was not itself arranged or performed by the manipulated person (Mele, 1995, p. 166).

So for Mele (1995), certain ways of acquiring preferences rule out responsibility. If his view is correct, control-bypassing interventions would be illegitimate for the same reason as other responsibility-thwarting interventions are. However, we shall remain uncommitted on the responsibility question[43] and instead analyze their relevance for our present concern. The debate over differences between acquiring preferences via natural and societal forces or manipulations by nefarious neuroscientists pertains directly to Levy's (2007) parity claim. Recall his parallel distinction: Some interventions *"bypass the agent's (rational) capacities altogether"*. Unlike Mele (in regard to responsibility), Levy denies the normative relevant difference of control bypassing interventions. Let us try to get a firmer understanding of what bypassing interventions are and whether they are normatively different.

Direct vs. indirect interventions

The thousand ways to change another's mind fall on a popular view into two camps: direct and indirect interventions. Direct interventions are those that work 'directly on the brain' such as DBS, TMS and psychoactive substances, and might bypass control whereas indirect interventions are somehow more remote. Sometimes the difference is cast in terms of interventions into the internal or external world. Strictly speaking, all these distinctions collapse: Any change in the external world must, in order to effectively change minds, cause internal (neuronal) changes; every indirect intervention affects brain activity. And most internal interventions which supposedly work directly on the brain (e.g. TMS stimuli) also pass through external space and the body. However, although these distinctions are not strictly mutually exclusive, I think they can be reconstructed for normative purposes under the guiding idea of self-determination.

I suggest considering as indirect interventions those stimuli which have to be perceived sensually

[43] It should be noted that existing legal systems are quite reluctant to accept manipulation or 'brainwashing' defences. This may change in light of neurotechnological mind interventions. Nevertheless, manipulated agents cannot be granted a carte blanche to violate norms as they please and should not be exonerated automatically (Bublitz & Merkel, 2009).

(hear, smell, see, feel) and processed psychologically including communication through word and sound, but also images or smells. In contrast, direct interventions are stimuli that reach the brain on other routes than perception, 'working directly on the brain', although from stimulus to brain change various metabolic processes might be involved. Pharmaceutically induced changes in the level of neurotransmitters in synaptic clefts or changes in electromagnetic fields in brain areas are primarily not psychological but brain processes, i.e. chemical or physical reactions. They follow the laws of nature whereas changes induced by perceptions of the world somehow relate to what is being perceived, to the psychological setup of the perceiver and follow 'psychological laws', even though they are realised by brain activity.[44] Again, these are not strict dichotomies: Indirect interventions involve light rays or sound waves carrying information and hence work, in a sense, also directly on the brain. Psychological processes are not only involved in processing information of what is being perceived, but also in processing direct interventions. Attaining new states of mind through pharmaceuticals surely involves some psychological processes at same stage – where and how, we do not know. This is the unsolved mind-brain mystery, but I assume that a person's psychological constitution significantly affects the outcome of an intervention and accounts for their interpersonal differences.

These objections notwithstanding, in terms of control the rough distinction between direct and indirect is tenable for most interventions. Individuals have most control over interventions they perceive especially when they rise to the level of conscious awareness. A further distinction can be drawn between consciously und subconsciously processed interventions. According to functional accounts of consciousness such as the global workspace model, contents in consciousness are broadcasted and received by specialised submodules. Consciousness is understood as a global information exchange enabling 'communication' between different parts of the brain and facilitating higher cognitive functions as working memory (Baars, 1997). If this is correct, conscious phenomena are better controllable than subconsciously processed interventions, control over the latter is much more limited. But even subconsciously processed information involves psychological processes. Before preferences or moods are transformed, stimuli are presumably 'checked' against or aligned with existing moods, preferences, etc. So, stimuli working their way through conscious awareness are more controllable, whereas subconsciously processed stimuli are less controllable.

Direct interventions are qualitatively different, presumably bypassing these psychological (not necessarily rational) processes altogether. Roughly one could say that indirect interventions are inputs into

[44] Speaking of psychological processes is, admittedly, vague. Tentatively, I suggest that psychological processes are those which can be best described by reference to psychological properties or mental states of persons such as fear or excitement instead of neuronal or physical occurrences. The distinction between indirect and direct interventions does not rely on commitment to a particular mind-brain theory. Presumably, the differences between causal pathways into the mind could be reformulated in reductionist terms without losing their peculiarities on which the normative differences are based, even if 'psychological laws' can be fully reduced to physical laws.

the cognitive machinery our minds are somehow adapted to process, whereas direct interventions change the cognitive machinery itself. Direct interventions somehow strain the connection between the individual and the world, be it appropriate or not. Admittedly, at some point this distinction might become unsustainable since indirect interventions could also change the machinery or the personality supervening on it. Yet, these changes are somehow more in line or accordance with the existing personality structure and preserve the individual's authenticity.[45] However, it should be noted that some indirect interventions, especially olfactory stimuli, are much less controllable than others, demonstrated by the special way smells sometimes evoke memories and feelings.[46] Fortunately, the normative dualism of intervention that I would like to propose does not primarily rely on the direct/indirect distinction. It only provides a first orientation for normative assessments based on the notion of control over ones mental sphere. And in this regard, even if interventions differ only gradually, they may differ significantly.

Awareness & evasion

Interventions may also differ in other control-related aspects. For instance, persons have greater powers to counteract interventions they are aware of. Communication seems effective only if the receiver gives the speaker a chance, an inner willingness to be persuaded. At least, one can revise arguments and use techniques as inner counterspeech, so that persons can at least drastically reduce, say, the impact of TV ads by reminding themselves of the manipulative business marketing is. In contrast, mental powers to counteract direct interventions are limited. Even if one is aware of being exposed to TMS or psychoactive substances, it is hard to withstand their effects through inner willing. Since means of resistance are limited, direct interventions can be considered more powerful.

Moreover, sometimes persons can evade exposure to stimuli. The sense organs are the gates to our minds, at least the eyes we can shut. So interventions carried out covertly are especially worrisome (e.g. odourless oxytocin) leaving no chance for escape or counteraction. Being unaware of stimuli, another control-undermining effect may set in, i.e. misattribution. When persons experience changes in emotional states they seek explanatory causes. Only those stimuli a person is aware of are candidates for attributing

[45] Of course, authenticity is one of the most cited and likewise challenged notions in the enhancement debate. But here it has a different normative function. In the enhancement debate it is often understood as an interest to be observed by oneself in self-transformations, here, much less problematic, it designates an interest to be protected against others. The contested issue of what an authentic personality consists of can be left to the individual.

[46] The olfactory system seems to be closely connected to emotional areas of the brain (Stockhorst & Pietrowsky, 2004). Olfactory stimuli illustrate the hairsplitting character of the direct/indirect distinction. The smell (and look) of fresh flowers creating sympathy in another is an indirect intervention whereas nasally administered cocaine is direct, although causal routes into the body might be identical.

causality. In empirical studies, arousal was induced in test persons through covert administration of drugs and, not surprisingly, subjects misattributed changes to other cues (Schachter & Singer, 1962). This is in a way facilitating self-deception and hence undermining self-control. Hence, sub specie control distinctions between interventions can be drawn by taking into account several control-related aspects – but are they normatively relevant?

Normative considerations

The right to mental SD commands respect by others. As we have seen, in terms of control some interventions are more worrisome than others, yet a normative standard is needed to evaluate which ones are illegitimate. For instance, one may claim that respect for mental SD obliges others to always use the least invasive means to achieve a given goal. This seems to be implied by Mele's (1995) bypass criterion. Strictly speaking, no control is bypassed in e.g. spraying oxytocin as we simply do not have control capacities over that causal route. Bypassing only makes sense in comparison: Route 2 bypasses an obstacle to be encountered on route 1. And so, there is a hidden normative assumption in Mele's (1995) criterion: If there is a causal route over which agents have more control, taking another route over which they have less control amounts to bypassing. This implies that using the route best controllable by affected persons is the default option for changing other people's minds. Tentatively, I would suggest this is a reasonable approach. At least, one could say that taking the less-controllable route is in need of strong justification. And this leads us to my main argument for normative dualism: the different normative status of interventions.

Normative status of interventions

Indirect interventions are often protected by other rights, whereas direct ones are usually not. As we have seen, indirect interventions are perceptible changes in the external environment. Oftentimes persons have legally protected interests to change the outside world. Property rights allow proprietors to change the appearance of objects and places, e.g. setting up a supermarket or painting their house as they wish even though this changes minds of perceivers. Most importantly, freedom of speech protects communication including audiovisual stimuli. Special rules apply to restricting speech (e.g. defamatory or deceptive speech) beyond which every act of communication is legally privileged. If communication was to be regulated only because it changes minds, social life would suddenly become mute and expressionless.

In contrast, direct interventions are in most cases not covered by special rights.[47] There simply is no

[47] Note that direct interventions may be justified by other reasons, e.g. (coerced) medication of mentally ill. This relates to the question whether interferences with the right to mental SD may be justified on other grounds. Here, we are

legally protected interest in spraying oxytocin or changing magnetic fields (of course, changing minds of others as such is not a legally protected interest). Therefore, no reasonably understood freedom is restricted if direct interventions were prohibited, but social life as we know it would collapse if indirect interventions were prohibited.

Still, it should be analyzed whether and which limitations might be placed on freedom of speech in virtue of their mental effects on others. Perhaps neuroscientific insights into decision-making and persuasion may shed new light on the cognitive capacities and susceptibility of listeners. However, in the end it is a normative decision and the ethics of rhetoric have been discussed since the days of the Sophists. Therefore, I would contend that every change in the minds of others caused by communication, and with that I mean the semantic content of a message, is legitimised by free speech. This demonstrates the difference between direct and indirect interventions. Curbing the latter immediately raises free speech issues and balancing rights, restricting the former does not. Dennett's (1984) incessantly preaching philosopher may therefore transform all beliefs and desires of another without raising legal suspicion. Note, however, that free speech does not place obligations on others to listen. If you happen to sit next to Dennett's philosopher on a flight to Australia, your right to mental SD (not to listen) must be balanced against the philosopher's freedom of speech. But these are special cases (called 'captioned audiences'). Generally, as a consequence of the normative difference between interventions, changing mental states – even harming – others through indirect interventions is permissible to a great extent while comparably less harmful outcomes via direct interventions interfere with mental SD.

Volenti non fit iniuria – consent

Finally, persons can consent to having their minds changed. We have only discussed cases without consent. Usually we do not wander through the world consenting to mind interventions. In a world full of utmost respect for mental SD, we would constantly ask each other kindly whether we are allowed to change their minds. Of course, this is ridiculous as they answer is often clearly affirmative. For instance, if a restaurant waiter declares that the paint on the walls and the flowers on the table have been carefully chosen to create a nice atmosphere and may together with the smells of fresh food emitting from the kitchen increase appetite, no one would bother. If he tells you that substances are used in cooking that do not do much for the taste but suppress feelings of overeating to sell more desserts, we would care. So there is a kind of seduction we want for which consent can be assumed without expression, and there is a kind of influence we despise. Only the latter is worrisome. But even if we do not want to be influenced by some interventions, everyone must be deemed as having consented to them because of his participating in an urban society in the informational age. The legal doctrine capturing this is called venire contra factum

---

establishing what may count as an interference with mental SD in the first place.

proprium ('estoppel' may be the English-law counterpart), which roughly states that the law does not need to recognise self-contradictory behaviour. Walking over Times Square watching billboards one cannot deny consenting to mind-change. This again relates to the normative differences in changing the external world. One cannot expect the world to be free from stimuli affecting our minds. Oftentimes, particular persons are entitled to design the world and we want the world to be designed, and this entails affecting perceivers. The outside world is something society must and wants to create meaningfully, and hence, one must be deemed as having consented to unavoidable exposure if one lives in a culturally rich society. More generally, one might say that as long as persons are aware of external, indirect stimuli, they consent to exposure unless they withdraw. Problems arise only when persons are not aware of or cannot escape stimuli.

Framework of illegitimate interventions

After all, we can weave the ends together to formulate some elements of a legal framework of illegitimate mind interventions: The law should recognise a right to mental SD that has to be respected by other persons. However, as social life is constant interaction between minds, not any stimuli changing other people's minds can be considered illegitimate or in need of justification. Instead, several factors have to be taken into account.

First, the intervention must be considered as having undermined mental SD in a legally relevant sense. The new mental states produced by the intervention have to have some relevance for thought processes, will-formation, opinions, preferences or other elements of one's personality, emotional well-being or mental health (these categories overlap). Regarding emotions, their appropriateness is an important issue. The more intense the interference, the more likely its illegitimacy, particularly worrisome are interventions undermining capacities of conscious self-control, strength of will (inducing weakness of will)[48] or diminishing capacities on which legal autonomy of subjects is grounded. Most of the millions of visual or auditory stimuli entering our mind day by day are either trivial as they lack significance or easily sheddable and shall not be of further interest. The example interventions changing moods (e.g. DBS), moral preferences or appraisal of other persons and therewith thought-processes, will-formation (e.g. oxytocin and TMS) or opinions (philosopher) are strong enough to warrant further scrutiny.

Secondly, when interventions cause potentially worrisome new mental states, the way they have been brought about needs to be evaluated. The guiding idea is that over some interventions persons have more control than over others. Especially worrisome are covertly administered or direct interventions bypassing

[48] Philosophy has grappled with weakness of will for ages. Recently, psychological theories compared strength of will to strength of muscles. Willpower may be depleted (and strengthened) by repeated exercise. Then, manipulating strength of will seems to be a realistic possibility (Levy, 2007).

psychological control capacities.

Thirdly, the normative status of the interventions has to be taken into consideration. Some interventions, especially communicative, are themselves protected by strong rights whereas others, especially direct interventions, are not. Therefore, the permissibility of an intervention depends on its outcome, means and normative status. From this we can deduce a normative dualism of interventions: *Prima facie, indirect interventions are permissible, indirect ones illegitimate.* Of course, every intervention needs careful analysis of its own.

This normative dualism stands in some opposition to Levy's (2007) parity principle, even though the former only applies to the law, the latter to ethics. Nonetheless, it may be worth to recapitulate the normative differences opposing it: Why should it matter to someone who has been influenced without consent to, say, purchase something he does not really want, whether he has been exposed to a buying-increasing substance or the charms, in words and person, of a saleslady? Because he cannot expect the environment to be free from any elements affecting him. Even if the saleslady's charms are as 'irresistible' as the substance, she has a (personality) right to be as charming as she wishes which prevails over his mental SD. He can, however, expect respect for his mental SD when it comes to substances as there is no legally protected interest in emitting them. If unconvinced consider it from the intervener's perspective: Should we tell the saleslady not to be charming? There is a clear normative difference between ways of changing minds.

This framework has to be expanded to incorporate other factors. Up to now, only single interventions were considered. In real-life there is no single stimulus, but sets of stimuli (e.g. 1000 ads plastered on billboards and broadcasted on TV) and for them to be combined in a legal evaluation rules of imputation have to be applied. Special criteria have to be worked out for emotional injuries, causality and foreseeability. Moreover, at this stage the model is insufficient to cover group processes in which manipulation or 'brainwashing' might most likely occur. Furthermore, it does not apply to parent-child or other special relationships since guardians have *legal duties* to interfere with their children's minds (education). Still, I think that even there indirect/direct distinction has some relevancy. Last but not least, thresholds for seriousness have to be defined, requiring close collaboration with empirical sciences. And above all, even if interventions interfere with mental SD, the question arises whether they may be justified by prevailing interests: Is mental SD an absolute right? Can interests of others or the state or the common good (Vedder & Klaming, 2010) outweigh mental SD? These are open questions for the law to address and any answer has to be grounded on a framework of illegitimate interventions.

Case applications

What does the proposed framework have to say about the example interventions? DBS, TMS and pharmaceuticals altering mental states to a severe degree are direct interventions, hard to counteract and hence undermine individuals' mental SD. Unless there are other prevailing interests, they are illegitimate. Thus, covertly spraying oxytocin (if effective) violates mental SD. Note that this does not imply that interveners commit a criminal offence as not every unlawful action constitutes one. Rather, at least in codified-law jurisdictions, criminal statutes indicating the prohibited behaviour are necessary, and currently,

there are to my knowledge no criminal offences against the mind. For German Law, I have proposed to introduce such an offence (Bublitz, 2011; Merkel, 2009).

Dennett's preaching philosopher is not acting unlawfully, as his means of changing minds, i.e. through speech, is indirect, does not bypass control capacities and enjoys special legal protection. With the exception of special situations (captioned audiences), changing minds through speech does not infringe upon mental SD. Marketing is surely the hardest case as it requires balancing freedom of speech and mental SD. Roughly, as long as the persuasiveness of ads relies on the content of the message it is legitimate. Special marketing forms targeting subconscious mechanisms, however, are problematic. Perhaps the irony of neuromarketing lies in the fact that the more scientific valid data demonstrate that its effectiveness is due to other factors than the persuasiveness of the message, the more reasons there are for curbing it. Obviously, marketing and other indirect means of changing minds require very different normative considerations than direct interventions. Yet, finding appropriate ways of dealing with new neurotechnologies gives reason to reconsider traditional means of changing minds too.

To conclude, the normative principles for changing other people's minds are vague. Even in the age of neuroscience, neither persons nor their legal protection can and should be reduced to their bodies. Thus, recognising and refining a mind protecting right and delineating its limits is one of the main challenges for neurolaw.

References

Ariely, D., & Berns, G. (2010). Neuromarketing: the hope and hype of neuroimaging in business. *Nature Reviews Neuroscience, 11*, 284-292.

Baars, B. (1997). *In the theater of consciousness: The workspace of the mind.* Oxford: Oxford University Press.

Bernsdorff, M. (2006). Commentary on Art. 10. In J. Meyer (Ed.), *Charta der Grundrechte der Europäischen Union* (2nd edition, 187-195). Baden-Baden: Nomos.

Boire, R. (2003). On cognitive liberty. *Journal of Cognitive Liberties, 4 (1).* Retrieved from http://www.cognitiveliberty.org.

Borowsky, M. (2006). Commentary on Art. 3. In J. Meyer (Ed.), *Charta der Grundrechte der Europäischen Union* (2nd edition, 118-133). Baden-Baden: Nomos.

Brasil-Neto, J., Pascual-Leone, A., Valls-Sole, J., Cohen, L., & Hallet, M. (1992). Focal transcranial magnetic stimulation and response bias in a forced-choice task. *Journal Neurology, Neurosurgery, Psychiatry, 55,* 964-966.

Bublitz, C., & Merkel, R. (2009). Autonomy and authenticity of enhanced personality traits. *Bioethics, 23 (6)*, 360-374.

Bublitz, C. (2011). Der (straf-)rechtliche Schutz der Psyche. *Rechtswissenschaft*, 28-69.

Clayton, R., & Thomlinson, H. (2000). *The law of human rights.* Oxford: Oxford University Press.

Damasio, A., (1994). *Descartes' Error: emotion, reason, and the human brain*. New York: Putnam.

Dennett, D. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge: MIT Press.

Faber, H., (1977). *Die innere Geistesfreiheit und suggestive Beeinflussung*. Berlin: Duncker & Humblot.

Fehr, E., Baumgartner, T., Heinrichs, M., Vonlanthen, A., & Fischbacher, U., (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron, 58,* 639-650.

Galert, T., Bublitz, C., Heuser, I., Merkel, R., Repantis, D., Schöne-Seifert, B., & Talbot, D. (2009). Memorandum: Das optimierte Gehirn. *Gehirn & Geist*, 40-48.

Greely, H., Sahakian, B.,Harris, J., Kessler, R., Gazzaniga, M., Campbell, P., & Frah, M. (2008). Towards responsible use of cognitive-enhancing drugs by the healthy. *Nature, 456,* 702-705.

Haji I., & Cuypers S. (2001). Libertarian free will and CNC manipulation. *Dialectica, 55 (3),* 221-238.

Kane, R. (1998). *The significance of free will.* Oxford: Oxford University Press.

Kant, I. (1797). *Metaphysik der Sitten* [Metaphysics of Morals]. Academy Edition.

Kohno, T., Denning, T., & Matsuoka, Y., (2009). Security and privacy for neural devices. *Journal Neurosurgery Focus, 27*, 1-4.

Kolber, A. (2007). Pain detection and privacy of subjective experience. *American Journal of Law & Medicine, 33,* 433-449.

Kosfeld, M., Heinrichs, M., Zak, P., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435,* 673-676.

Levy, N., (2007). *Neuroethics*. Cambridge: Cambridge University Press.

Lillich, R. (1991). The Soering case. *American Journal of International Law, 85,* 128-150.

McInerney, P. (2009). Negligently inflicted psychological harm and the 'sudden shock' requirement: A comparative analysis. *Electronic Journal of Comparative Law, 13(3).* Retrieved from http://www.ejcl.org/133/abs133-6.html.

Mele, A. (2006). *Free will and luck*. Oxford: Oxford University Press.

Mele, A. (1995). *Autonomous agents.* Oxford: Oxford University Press.

Merkel, R. (2007). *Intervening in the brain: changing psyche and society*. Berlin: Springer.

Merkel, R. (2009). Die Verbesserung der condicio humana und ihre strafrechtlichen Grenzen. *Zeitschrift für die gesamte Strafrechtswissenschaft, 121*, 919-953.

Neziroglu, I. (2007). Comparative analysis of mental and psychological suffering as torture, inhuman or degrading treatment or punishment under International Human Rights Treaty Law. *Essex Human Rights Review, 4 (1),* 65-81.

Noordhof, P. (2003). Believe what you want. *Proceedings of the Aristotelian Society, 101,* 247-265.

Pauen, M. (2007). Neuroökonomie – Grundlagen und Grenzen. *Analyse & Kritik, 29*, 24–37.

Pereboom, D. (2003). *Living without free will.* Cambridge: Cambridge University Press.

Pettit, P., & Smith, M. (1996). Freedom of belief and desire. *The Journal of Philosophy, 9,* 429-449.

Rawls, J. (1971). *A theory of justice*. Cambridge: Harvard University Press.

Rengeling, H., & Szczekalla, P. (2004). *Grundrechte in der Europäischen Union*. Köln: Heymann.

Sandel, M. (2007). *The case against perfection: ethics in the age of genetic engineering*. Cambridge: Harvard University Press.

Schachter, S., & Singer, J. (1962). Cognitive, social and physiological determinants of emotional states. *Psychological Review, 69,* 379–399.

Schleim, S. (2010). *Die Neurogesellschaft*. Hannover: Heise.

Stockhorst, U., & Pietrowsky, R. (2004). Olfactory perception, communication, and the nose-to-brain pathway. *Physiology and Behaviour, 83*, 3-11.

Vedder, A., & Klaming, L. (2010). Human enhancement for the common good – using neurotechnologies to improve eyewitness memory. *American Journal of Bioethics Neuroscience, 1 (3)*, 22-33.

Vermeulen, B. (2006). Commentary on Art. 9. In P. van Dijk, F. van Hoof, A. van Rijn, & L. Zwaak (Eds.), *Theory and Practice of the European Convention on Human Rights* (4th edition, 751-771). Antwerpen: Intersentia Press.

Warren, S., & Brandeis, L. (1890). The right to privacy. *Harvard Law Review, 4 (5),* 193-220.

Wegner, D. (2005). Who is the controller of controlled processes? In R. R. Hassin (Ed.), *The new unconscious* (19-36). Oxford: Oxford University Press.

Young, L., Camprodon, J., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right

temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Science, 107(15),* 6753-6758.

Chapter 5
The influence of neuroscientific evidence on legal decision-making: the effect of presentation mode

Laura Klaming
Tilburg University
Tilburg Institute for Law, Technology, and Society
✉ l.klaming@uvt.nl

Abstract Findings from neuroscience research are increasingly advancing our understandings of the neural correlates of human behaviour, cognition and emotion. These findings are beginning to gain visibility in the legal system, including the courtroom. To an increasing extent, judges are being confronted with neuroscientific evidence concerning the degree to which the suspect should be held criminally responsible, the likelihood of future offending and the presence of emotional pain. By directly measuring brain activity, neurotechnologies hold the promise of increasing the quality of evidence in legal proceedings, and could therefore be of great value to the legal system. However, such practice has important implications including the possibility that neuroscientific evidence is overly persuasive, thereby unduly affecting legal decision-making. The presentation of visual information, i.e. brain images, may even increase this effect. In fact, several researchers have recently raised the concern that neuroscientific evidence, especially brain images, may be perceived in court without sufficient critical appraisal and should therefore be inadmissible in court. However, there is currently limited empirical support for this claim, which raises the risk of drawing premature and invalid conclusions regarding the responsible use of neuroscientific evidence in the courtroom. So far, the difference between the effect that verbal and visual neuroscientific evidence may have on legal decision-making has not been sufficiently taken into account. Additionally, the possibility that visual neuroscientific evidence, i.e. brain images, increases the comprehensibility of the information rather than contributes to the potentially overly persuasive effect of neuroscientific evidence has not yet been considered. The present paper deals with the effect that neuroscientific evidence may have on legal decision-making taking presentation mode into account.

Keywords neuroscientific evidence, legal decision-making, brain images, presentation mode

Introduction

In 2007, a 63-year-old man stabbed a friend nine times, as a result of which she deceased. The suspect declared that he was annoyed by the victim's behaviour. He furthermore declared that he saw that the victim had lost a lot of blood as a result of the stabbing and lost her consciousness several times. When she regained consciousness and tried to get up he stabbed her again. At the time of the incident, the suspect was intoxicated with alcohol and cocaine. According to one of the expert witnesses, a behavioural neurologist, the suspect's behaviour during the incident was affected by damage to his frontal lobes. More specifically, the brain damage had rendered the suspect unable to control his impulses and reflect on his

actions in difficult situations. The alcohol and cocaine were believed to have aggravated his impulsive behaviour. Additionally, the expert witness stated that the suspect's brain damage had interfered with his free will. On the basis of inter alia the neurologist's testimony, the presiding judge decided that the suspect had acted intentionally. More specifically, he stated that although the suspect's behaviour was affected by the frontal lobe damage, he did not lack complete insight into the consequences of his actions. According to the judge, the suspect was aware of the possibility that the victim would die as a result of the harm that he was inflicting on her. Consequently, the court decided that the suspect had committed the act of intentionally killing someone (manslaughter). The judge also considered the neurologist's testimony along with the testimony of a psychiatrist and a psychologist when deciding about the suspect's responsibility. According to the neurologist, the suspect was severely diminished responsible due to his frontal lobe damage, which as stated above had interfered with his free will. The presiding judge accepted this explanation and adopted the expert witness' diagnosis of severely diminished responsibility, which eventually resulted in reduced sentencing of 18 months imprisonment, plus detention during Her Majesty's pleasure (District Court Amsterdam, 2008)[49].

This case clearly demonstrates the potential of using neurotechnologies in a legal context. Besides the determination of intentionality and the assessment of responsibility (Aharoni, Funk, Sinnott-Armstrong, & Gazzaniga, 2008; Gazzaniga, 2008; Greene, & Cohen, 2004; Vincent, 2008), neurotechnologies have numerous other legitimate legal applications, both in criminal and civil law. These include determining the likelihood of future offending and the treatment of criminal behaviour (Gazzaniga, 2008; Greely, 2008). According to Greely (2008), neuroscience may provide us with more effective rehabilitation through treatment aiming at directly changing our brains. Chemical castration for sex offenders, i.e. a pharmacological treatment aimed at a decrease in testosterone levels and resulting decrease in sexual thoughts and behaviour, is an example of a neuroscience-based treatment of criminal behaviour. Additionally, it has been suggested that neurotechnologies might play a role as a method to detect deception (Kozel, Johnson, Mu, Grenesko, Laken, & George, 2005; Langleben et al., 2002; Spence, Farrow, Herford, Wilkinson, Zheng, & Woodruff, 2001) or to enhance eyewitness memory (Klaming & Vedder; 2009; Vedder & Klaming, 2010). Research has for instance indicated that certain brain areas that are associated with high-level executive functions including areas in the frontal cortex are more active during deception, which has been argued to be due to the increased cognitive effort involved in lying (Langleben et al. 2002; Spence et al.

[49] Information about this case was extracted from www.rechtspraak.nl, the website of the Dutch Judiciary and the Supreme Court of the Netherlands. Since only summaries of court cases can be found on the website, it is unclear how the neurologist examined the suspect and exactly what kind of evidence he presented to the court besides the written report of his examination of the suspect.

2004). As truth-finding is the core goal of criminal proceedings, accurate and reliable lie-detection would be of great value to the criminal justice system. Besides their potential applications in criminal law, neurotechnologies also have potentially valuable applications in the civil justice system. In civil law, neurotechnologies might be used for pain detection or to determine mental competence (Bremner, 2007; Grey, 2007; Jones, Buckholtz, Schall, & Marois, 2009; Kolber, 2007; Ochsner et al., 2006; Peyron, Laurent, & Garcia-Larrea, 2000; Tovino, 2007). Several studies have used neuroimaging technologies to explore the neural correlates of physical and emotional pain (Bremner, 2007; Ochsner et al., 2006; Peyron et al., 2000). Studies on the neural correlates of posttraumatic stress disorder have for instance shown that dysfunction of the medial prefrontal cortex, hippocampus, and amygdala may underlie symptoms of the disorder (Bremner, 1999, 2007). More objective methods of pain proof in tort cases would certainly be of great value to the civil justice system.

Since neurotechnologies allow researchers to directly measure brain activity, they hold the promise of being more objective and accurate methods for the ascertainment of cognitive or affective states. Despite its potential value, it is however important to note that the probative value of neuroscientific evidence is yet limited as a result of insufficient scientific proof that at this point in time neurotechnologies can be employed to accurately and reliably determine criminal responsibility, predict the likelihood of future criminal behaviour, detect deception or detect physical and emotional pain. Although there are some studies that have shown that neurotechnologies could be used for these purposes, more scientific research is necessary before any valid conclusions about the accuracy and reliability of using neurotechnologies for the detection of deception, the assessment of criminal responsibility, or pain detection can be drawn. Nevertheless, as the summary of the case described above demonstrates, neuroscientific evidence has already entered the courtroom in the Netherlands. Consequently, neuroscientific evidence has already influenced legal decisions and will most probably be cited in courts more frequently in the coming years as technologies advance.

Putting aside the fact that the use of neurotechnologies for legal purposes is currently questionable due to the limited scientific research, another important challenge of applying neurotechnologies in a legal context refers to the possibility that neuroscientific evidence is inappropriately persuasive and may therefore unduly affect legal decision-making. Given that neuroscientific evidence has already influenced legal decisions, this may in fact be one of the most important challenges at this point in time. In fact, several researchers have already claimed that neuroscientific evidence may be perceived without sufficient critical appraisal by legal decision-makers and should therefore be inadmissible in court[50] (Dumit, 2004; Jelicic & Merckelbach, 2007; Reeves, Mills, Billick, & Brodie, 2003; Sinnott-Armstrong, Roskies, Brown, & Murphy,

[50] It is important to note that the debate about the responsible use of neuroscientific evidence in court has mainly focused on the use of neurotechnologies for the assessment of responsibility.

2008; Garland & Glimcher, 2006). Since empirical research supporting this claim is yet very scarce – as some researchers acknowledge (e.g. Sinnott-Armstrong et al., 2008) – it is impossible to draw a valid conclusion about the responsible use of neuroscientific evidence in the courtroom at this point in time. Moreover, it is crucial to differentiate between the verbal and visual presentation of neuroscientific evidence when discussing the effect of neuroscientific evidence on legal decision-making. This differentiation is important because verbal neuroscientific evidence, i.e. a written and/or oral explanation of the expert witness' assessment, may have a different effect on legal decision-making than including brain images with the expert testimony. More specifically, three distinct scenarios are possible. Firstly, it is possible that both verbal and visual presentations of neuroscientific evidence are overly influential and unduly affect legal decision-making. Secondly, it is possible that neither verbal nor visual presentations of neuroscientific evidence are overly persuasive, at least not more persuasive than other types of expert testimony. And thirdly, it is possible that one of the two, but not the other induces a bias on legal decision-making. Obviously, knowing how both verbal and visual neuroscientific evidence affect legal decision-making is necessary in order to make an informed and valid decision about the admissibility of neuroscientific evidence in court. So far, discussions on the admissibility of evidence based on neuroscience has not considered the possibility that the verbal presentation of neuroscientific evidence may have a different effect on legal decision-making than the additional presentation of visual neuroscientific evidence, i.e. brain images that support the verbal explanations of the expert witness. The aim of the present paper is to explain how neuroscientific evidence may affect legal decision-making taking presentation mode into account.

The perils of using neuroscientific evidence in the courtroom

The implications of using neurotechnologies for legal purposes continue to be discussed (e.g. Farah, 2002; Feigenson, 2006; Gazzaniga, 2005; Glannon, 2005; Greely, & Illes, 2007; Jelicic & Merckelbach, 2007; Morse, 2006; Reeves et al., 2003; Sinnott-Armstrong et al., 2008; Wolpe, Foster, & Langleben, 2005). One of the major concerns already briefly mentioned above refers to the fact that research on the use of neurotechnologies for legal purposes such as the assessment of intentionality or responsibility or the detection of deception is still in its infancy. As a consequence, we cannot yet draw reliable and valid conclusions about a suspect's intentionality or responsibility or about the truthfulness of his statements based on his brain activity (Jelicic & Merckelbach, 2007; Morse, 2006). Another concern of using neurotechnologies both more generally and for legal purposes refers to privacy (Farah, 2002; Wolpe et al., 2005). Neuroimaging data can reveal a great deal of information about an individual, such as aspects of personality, mental illness or predisposition to drug addiction (Canli & Amin, 2002; Childress et al., 1999). When using fMRI for legal purposes, this kind of information could also be revealed which raises the question who should have access to the data obtained by neurotechnologies. Obviously, privacy of the mind in relation to the appropriate use of information gained by exploring an individual's brain processes in criminal (and civil) settings needs to be discussed – and potentially be regulated – before neurotechnologies can be used for legal purposes.

In addition to these and other concerns, one of the most important challenges of applying

neurotechnologies for legal purposes refers to the possibility that evidence derived from neuroscience might be accepted without much critical appraisal and may therefore unduly affect legal decision-making (Dumit, 2004; Jelicic & Merckelbach, 2007; Reeves et al., 2003; Sinnott-Armstrong et al., 2008; Garland & Glimcher, 2006). This assumption is supported by recent research demonstrating that people view explanations of psychological phenomena as more believable if these explanations contain a neuropsychological component (Weisberg, Keil, Goodstein, Rawson, & Gray, 2008). Weisberg and her colleagues (2008) demonstrated in a series of experiments that including neuroscience information in explanations of psychological phenomena had an influence on participants' satisfaction with these explanations. Subjects were asked to read short descriptions of psychological phenomena and short explanations for these phenomena, and were then instructed to rate how satisfying they found the explanations. The results indicated that participants who had received explanations that contained neuroscience information were more satisfied with the explanations of the psychological phenomena as compared to participants who had received explanations that did not include neuroscience information. The effect was found to be particularly strong for judgments of bad explanations of these phenomena, i.e. explanations that were circular restatements of the explained phenomenon. The effect of the *"seductive allure of neuroscience explanations"* (Weisberg et al., 2008, p. 470) was limited to non-experts, which suggests that training has a benefit on judgments of explanations (Weisberg et al., 2008). A similar effect with regard to the influence of neuroscience information was found in another study that explored the effect of brain images on judgements of scientific reasoning (McCabe & Castel, 2007). In this study, participants read short articles, for instance about the finding that watching TV is related to math ability. Depending on the experimental condition they were assigned to, participants were then presented with brain images, for instance brain images that showed that watching TV activated similar brain areas as completing arithmetic problems. Including visual information, i.e. brain images, with summaries of cognitive neuroscience data significantly increased people's ratings of scientific reasoning for those summaries as compared to no additional visual information and bar graphs. The visual information that was added to the explanations was completely redundant with the text that participants received. The finding that including brain images in the explanations of scientific findings increased people's ratings of scientific reasoning for those explanations suggests that visual neuroscience information, such as brain images, may have a particularly strong effect on evaluations of explanations, at least within the context of perceptions of cognitive neuroscience research. According to McCabe and Castel (2007) this effect is due to the fact that brain images are physical and therefore less abstract representations of cognitive processes, appealing to *"people's natural affinity for reductionist explanations"* (McCabe & Castel, 2007, p. 344).

Although these two studies analyzed the influence of neuroscience explanations on the public's perception of scientific research and not within a legal context, a similar effect might occur if neuroscience information is presented in court. It is important to note that the first study (Weisberg et al., 2008) explored the influence of verbal neuroscience information, i.e. a written description of a phenomenon, while the second study (McCabe & Castel, 2007) explored the influence of visual neuroscience information, i.e. a brainscan. It is important to make this distinction since both types of presentation are likely to have a different effect on people's judgments. Nevertheless, both studies found that neuroscience information has

an overly persuasive effect on judgments of scientific reasoning and there is therefore reason to assume that a similar effect may occur if neuroscience enters the courtroom. One recent study empirically supports the concern that neuroscientific evidence unduly affects legal decision-making by showing that students were more likely to find a hypothetical offender not guilty by reason of insanity if he had some kind of brain damage as presented in a brain image (Gurley & Marcus, 2008). In this experimental study, participants were presented with a summary of a criminal case and expert testimony regarding the suspect's mental condition and then had to render a verdict of either guilty or not guilty by reason of insanity. Defendants diagnosed with a psychotic disorder, defendants who had a brain lesion as demonstrated in a brainscan, and defendants who had a history of brain injury were more likely to be found not guilty by reason of insanity than those defendants who did not present any psychological or neurological testimony. A combination of a psychotic disorder, brain damage that had caused the disorder and neuroimaging evidence depicting the damage increased the likelihood of a not guilty by reason of insanity verdict as compared to either of these types of testimony alone. Although these findings are very interesting and clearly demonstrate that psychological and neurological evidence affect legal decisions, there is an important concern that warrants closer examination and that limits the possibility of drawing valid conclusions about the influence of neuroscientific evidence on legal decision-making more generally. This concern has to do with the fact that it is currently unclear what exactly the effect of presentation mode is, which complicates discussions about the admissibility of neuroscientific evidence in the courtroom[51].

Presentation mode

The concerns of researchers who have argued that neuroscientific evidence should be inadmissible in court because it is perceived as overly persuasive almost exclusively focus on the use of brain images, i.e.

[51] Besides the potential effect of presentation mode, there are additional issues that need closer examination before it will be possible to draw valid conclusions about the admissibility of neuroscientific evidence more generally. It is difficult to draw more general conclusions on the basis of Gurley and Marcus (2008) study, because the application of neurotechnologies to questions of responsibility is just one of the numerous legal applications of neurotechnologies. Consequently, knowing that neuroimages increase the likelihood of accepting an insanity defense does not tell us anything about the impact of neuroscientific evidence in general on other types of legal decisions. Additionally, responsibility is conceptualised differently in distinct jurisdictions, which is likely to affect judgments of responsibility. In the Netherlands for instance, a 5-point scale ranging from complete responsibility, slightly diminished responsibility, diminished responsibility, severely diminished responsibility to complete absence of responsibility is used in forensic practice (Barendregt, Muller, Nijman, & de Beurs, 2008), whereas in other countries determination of a suspect's responsibility is often a binary decision.

the visual presentation of information. It has been argued that judges and juries might get distracted by the 'pretty lights' in brain images resulting in reduced attention to an expert's explanation of the brain image, which has been referred to as *"Christmas tree phenomenon"* (Mobbs et al., 2007, p. 0699). If evidence concerning the intentionality, responsibility or truthfulness of a suspect is presented together with images of the suspect's brain, the vividness and the technological sophistication of these images may be very compelling and may even distract the judge's or jurors' attention away from the expert's explanation. Seeing a brain image of the suspect may easily push aside any concerns about the reliability or validity of neurotechnologies for the purpose of determining the suspect's intentionality, responsibility or truthfulness (Dumit, 2004; Henson, 2005; Mobbs et al., 2007; Reeves et al., 2003).

An additional problem with brain images besides their potential effect on attention has been outlined by Henson (2005), who argues that non-experts are likely to believe that it is possible to *"directly observe psychological constructs"* (Henson, 2005, p. 228) when they are presented with brain images. Judges and jurors might not be aware that brain scans present interpretative rather than 'photographical' images of what is occurring in the brain (Feigenson, 2006; Jones et al., 2009; Henson, 2005; Roskies, 2007). Hence, if they are presented with a brainscan that depicts some kind of dysfunction or damage they may easily believe that this malfunction caused the criminal behaviour without realizing that there may not be a causal link between the abnormality and the suspect's behaviour. Laypeople may have difficulty differentiating between the brain abnormality being proof of the dysfunctional behaviour and it being merely consistent with the dysfunctional behaviour.

Besides reduced attention and naïve realism, neuroimages may increase people's propensity to commit the fundamental attribution error, the tendency to overvalue dispositional and undervalue situational explanations (Ross, 1977). Being presented with a brainscan of a suspect may therefore prompt people to believe that the behaviour was caused by the brain damage and not by other factors when in fact behaviour is generally brought about by a multitude of factors including dispositional and situational factors.

Besides these arguments, some support for the concern that brain images are overly persuasive comes from research on legal decision-making. Studies have found that the visual presentation of information, including photographs and videotapes, is more persuasive than verbal descriptions of the same evidence (Bright & Goodman-Delahunty, 2006; Kassin & Garfield, 1991). For instance, in one study, participants were presented with either only verbal or verbal and photographic evidence of a crime scene and the victim's wounds and were asked to indicate whether they found the suspect guilty. The results of the study revealed that participants who had received photographs of the crime scene and the victim's wounds were more likely to find the suspect guilty than suspects who had not received photographic evidence (Bright & Goodman-Delahunty, 2006). This explanation is supported by the 'vividness effect', the idea that information has a greater impact on social judgment when it is emotionally interesting, detailed, and highly imaginable (Bell & Loftus, 1989; Nisbett & Ross, 1980; Reyes, Thompson, & Bower, 1980). According to Bell and Loftus (1989), vivid information may carry more weight in judgments because it may attract more attention, recruit more additional information from memory, be more available in memory, have a greater affective impact and be perceived as having a more credible source.

On the basis of these arguments, it seems obvious that neuroscientific evidence should be inadmissible in court. However, this conclusion would be premature for two reasons. Firstly, it is possible that only the visual presentation, but not the verbal presentation of neuroscientific evidence is overly influential. As previously mentioned, the debate on the admissibility of neuroscientific evidence has almost exclusively focused on brain images. Since no empirical research directly comparing the effect of verbal and visual neuroscientific evidence on legal decision-making is yet available, we do not know how persuasive verbal neuroscientific evidence is and we can therefore not draw valid conclusions about the responsible use of neuroscientific evidence in the courtroom at this point in time. In this regard, it is moreover important to note that verbal neuroscientific evidence may be as persuasive as other types of expert testimony, such as for instance DNA evidence, because neuroscientific evidence may be more tangible proof of a disorder or specific cognitive or affective state as compared to the testimony of a psychologist or a psychiatrist. Hence, we need empirical research that compares the influence of distinct types of expert evidence in order to make an informed and justified decision about the admissibility of neuroscientific evidence. Secondly, while the visual presentation of neuroscientific evidence may have a negative effect on legal decision-making, because it draws away attention from the expert explanation, is easily misinterpreted or is simply perceived as more credible, it is equally possible that the visual presentation of neuroscientific evidence has a positive effect on legal decision-making. Instead of confusing the decision-maker, presenting evidence visually may increase the comprehensibility of the information. The potentially positive effect of visual neuroscientific evidence – which has not yet been considered in debates on the admissibility of neuroscientific evidence in the courtroom – will be discussed in more detail in the following section.

The potentially positive effect of brain images

It is possible that instead of having a negative effect on legal decision-making, the visual presentation of neuroscientific evidence has a positive effect on legal decision-making, because it increases the comprehensibility of the expert witness' explanations. This is why regardless of the relevance of the findings of Gurley and Marcus (2008), it is important to consider alternative explanations for the finding that defendants who presented visual evidence for their brain damage were more likely to be found not guilty by reason of insanity, before drawing any conclusions about the responsible use of neuroscientific evidence in the courtroom. Although the findings of Gurley and Marcus (2008) suggest that jurors endow brain images with greater credibility, it is possible that the distorting effect of neuroimages was the result of complex information being presented visually rather than the brainscans per se. In contrast to what several researchers claim – i.e. that brain images confuse the decision-maker – brain images may in fact increase the decision-maker's ability to comprehend the information presented. This explanation is supported by dual coding theories. According to dual coding theories, people use two separate channels for the processing of information, one for auditory and one for visual information. Each channel is limited in the amount of information it can process (Clark & Paivio, 1991; Paivio, 1971). Presenting information both verbally and visually will therefore reduce the cognitive load on each of the channels and contribute to better processing of the information. Interesting research in this regard has been conducted by Hewson and Goodman-

Delahunty (2008), who found that mock jurors were more likely to comprehend DNA evidence when it was accompanied by multimedia instructions consisting of a narrated video sequence using both animation and textual content to explain DNA evidence. As they state *"these results are encouraging for criminal justice practitioners and courts that have been cautious in adopting innovative technologies or concerned that multimedia will tend to persuade and possibly mislead jurors rather than facilitate their understanding of complex expert evidence"* (Hewson & Goodman-Delahunty, 2008, p. 62). Other research supports this assumption (Brewer, Harvey, & Semmler, 2004; Kassin & Dunn, 1997; Morell, 1998). For instance, Morell (1998) found that people who received expert testimony in combination with a computer animation recalled information more accurately and in more detail than participants who received expert testimony without visual aids.

Presenting evidence visually may decrease the cognitive efforts required to comprehend the information, which may contribute to a better understanding of the evidence. Combining verbal and visual presentations of evidence may be the most effective means of communicating complex information to a layperson who subsequently has to make a decision on the basis of this information. Consequently, the visual presentation of neuroscientific evidence may have a positive rather than a negative effect on legal decision-making. If this assumption turns out to be true, visual neuroscientific evidence, i.e. brain images, should be admissible in court. Nevertheless, more empirical research on the influence of both verbal and visual neuroscientific evidence is beneficial. This research should explore whether, consistent with dual-coding theory, the visual presentation of neuroscientific evidence improves the comprehensibility of the information. Additionally, as mentioned above, it is essential that the influence of neuroscientific evidence is compared to the influence of other types of expert evidence such as for instance DNA evidence before any conclusions about the persuasiveness of neuroscientific evidence can be drawn.

Discussion and conclusion

The use of neurotechnologies in a legal context is promising; especially since directly measuring brain activity may yield more accurate and reliable evidence. Rapid developments and advances require discussing the possible persuasiveness of neuroscientific evidence. As soon as neuroscientific evidence is used in court more frequently, it will be necessary to know whether and to what degree it is overly influential and unduly affects legal decision-making. The debate concerning the admissibility of neuroscientific evidence in court has already begun with the vast majority of researchers claiming that neuroscientific evidence should be inadmissible because it is accepted without sufficient critical appraisal. These concerns have however exclusively focused on the visual presentation of neuroscientific evidence, i.e. brain images. Not knowing whether the verbal presentation of neuroscientific evidence has the same effect on legal decision-making as the visual presentation of the same evidence complicates the discussion about the responsible use of neuroscientific evidence.

Moreover, there is generally a risk of drawing premature and invalid conclusions without sufficient empirical support for the claim that neuroscientific evidence should be inadmissible in court because it is overly persuasive. An example of how opinions rather than empirical data have shaped the public discussion

in the past is the 'CSI effect' – the claim that jury members who are avid watchers of the television series Crime Scene Investigations have an unrealistic belief in the infallibility of forensic science (Cole & Dioso-Villa, 2007; Tyler, 2006). The CSI effect has received an enormous amount of media attention, which not only unsettled the public, but moreover resulted in calls for changes in the legal system. These calls were unsubstantiated because they lacked empirical support. Without empirical data, it is possible to plausibly argue for the opposite effect. With regard to the CSI effect, it is equally plausible that CSI watchers are more critical of scientific evidence because they have greater expectations about scientific evidence than can actually be delivered. In fact, recent empirical research supports this hypothesis (Schweitzer & Saks, 2007; Shelton, Kim, & Barak, 2006).

A similar effect may occur if the discussion on the responsible use of neuroscientific evidence is based on opinions rather than empirical data. The risk for unsubstantiated conclusions seems to be particularly high for visual neuroscientific evidence since the vast majority of the debate has focused on neuroimaging technologies. As outlined above, arguing that brain images should be inadmissible in court because they are overly influential is risky, because it is possible to plausibly argue that brain images reduce cognitive demands because they are less abstract representations of cognitive processes. Consequently, neuroimages may increase the comprehensibility of expert testimony rather than confuse the decision-maker. As described above, research on the visual presentation of expert testimony suggests that this assumption may be true (Brewer, Harvey, & Semmler, 2004; Hewson & Goodman-Delahunty, 2008; Kassin & Dunn, 1997; Morell, 1998). Hence, more research on the effect of presentation mode within the specific context of neuroscientific evidence is beneficial and would contribute to a more substantiated discussion about the responsible use of neuroscientific evidence in criminal (and civil) proceedings

If neuroscientific evidence turns out to be accompanied by an exaggerated belief in its infallibility and induces a bias on legal decision-making, regulatory efforts to mitigate this effect might be necessary. As mentioned above, studies on the influence of neuroscience information on judgments of scientific reasoning (Weisberg et al., 2008) have shown that the persuasive effect was limited to non-experts, which suggests that training has a benefit on judgments of explanations. Hence, training for laypeople, including judges, prosecutors and lawyers, about what can and what cannot be inferred from results derived by neurotechnologies may prove useful in order to prevent misinterpretation. Since neuroscientific evidence has already influenced legal decisions, as the brief description of the case in the introduction demonstrates, it is important to further discuss the use of neuroscientific evidence in the courtroom. It seems important to realise that some of the problems encountered with the use of neurotechnologies in the courtroom may turn out to be very similar to the problems encountered with other clinical methods and technologies.

References

Aharoni, E., Funk, C., Sinnott-Armstrong, W., & Gazzaniga, M. (2008). Can neurological evidence help courts assess criminal responsibility? Lessons from law and neuroscience. *Annals of the New York Academy of Sciences, 1124*, 145-160.

Barendregt, M., Muller, E., Nijman, H., & de Beurs, E. (2008). Factors associated with experts' opinions regarding criminal responsibility in the Netherlands. Behavioral Sciences and the Law, 26, 619-631.

Bell, B. E., & Loftus, E. F. (1989). Trivial persuasion in the courtroom: the power of (a few) minor details. *Journal of Personality and Social Psychology, 56 (5)*, 669-679.

Bremner, J. D. (1999). Alterations in brain structure and function associated with post-traumatic stress disorder. *Seminars in Clinical Neuropsychiatry,* 4(4), 249-55.

Bremner, J. D. (2007). Functional neuroimaging in post-traumatic stress disorder. *Expert Review of Neurotherapeutics, 7(4),* 393-405.

Brewer, N., Harvey, S., & Semmler, C. (2004). Improving comprehension of jury instructions with audio-visual presentation. *Applied Cognitive Psychology, 18*, 765-776.

Bright, D. A., & Goodman-Delahunty, J. (2006). Gruesome evidence and emotion: anger, blame, and jury decision-making. *Law and Human Behavior, 30*, 183-202.

Canli, T., & Amin, Z. (2002). Neuroimaging of emotion and personality: scientific evidence and ethical considerations. *Brain and Cognition, 50*, 414–431.

Childress, A. R., David Mozley, D., McElgin, W., Fitzgerald, J., Reivich, M., & O'Brien, C. P. (1999). *Limbic Activation During Cue-Induced Cocaine Craving. The American Journal of Psychiatry, 156 (1)*, 11-18.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review, 3 (3)*, 149-210.

Cole, S. A. & Dioso-Villa, R (2007). CSI and its Effects: Media, Juries, and the Burden of Proof. *New England Law Review, 41 (3)*, 435-470.

District Court Amsterdam (2008). LJN: BC9296.

Dumit, J. (2004). *Picturing Personhood*. Princeton, NJ: Princeton University Press.

Farah, M. J. (2002). Emerging ethical issues in neuroscience. *Nature Neuroscience, 5 (11)*, 1123-1129.

Feigenson, N. (2006). Brain imaging and courtroom evidence: on the admissibility and persuasiveness of fMRI. *International Journal of Law in Context, 2 (3)*, 233-255.

Garland, B., & Glimcher, P. W. (2006). Cognitive neuroscience and the law. *Current Opinion in Neurobiology, 16*, 130-134.

Gazzaniga, M.S. (2005). *The ethical brain*. New York: Dana Press.

Gazzaniga, M. (2008). The law and neuroscience. *Neuron, 60 (6)*, 412-415.

Glannon, W. (2005). Neurobiology, Neuroimaging, and Free Will. Midwest Studies in Philosophy, 29, 68-82Greely, H. T. (2008). Neuroscience and criminal justice: not responsibility but treatment. *Kansas Law Review, 56*, 1103-1138.

Greely, H. T., & Illes, J. (2007). Neuroscience-Based Lie Detection: The Urgent Need for Regulation. *American Journal of Law & Medicine, 33*, 377-431.

Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 359(1451)*, 1775-1785.

Grey, B. J. (2007). Neuroscience, emotional harm, and emotional distress tort claims. *American Journal of Bioethics, 7 (9)*, 65-67.

Gurley, J. R., & Marcus, D. K. (2008). The effects of neuroimaging and brain injury on insanity defenses. *Behavioral Sciences and the Law, 26*, 85-97.

Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? *The Quarterly Journal of Experimental Psychology, 58A (2)*, 193–233.

Hewson, L., & Goodman-Delahunty, J. (2008). Using multimedia to support jury understanding of DNA profiling evidence. *Australian Journal of Forensic Sciences, 40 (1)*, 55-64.

Jelicic, M., & Merckelbach, H. (2007). Hersenscans in de rechtzaal: oppassen geblazen! *Nederlands Juristenblad, 44*, 2794-2800.

Jones, O. D., Buckholtz, J. W., Schall, J. D., & Marois, R. (2009). Brain imaging for legal thinkers: a guide for the perplexed. *Stanford Technology Law Review, 5*. Retrieved from http://stlr.stanford.edu/pdf/jones-brain-imaging.pdf

Kassin, S. M., & Garfield, D. (1991). Blood and guts: general and trial specific effects of videotaped crime scenes on mock jurors. *Journal of Applied Social Psychology, 21*, 1459-1472.

Kassin, S. M., & Dunn, M. A. (1997). Computer-animated displays and the jury: facilitative and prejudicial effects. *Law and Human Behavior, 21 (3)*, 269-281.

Klaming, L., & Vedder, A. (2009). Brushing Up Our Memories: Can We Use Neurotechnologies to Improve Eyewitness Memory? *Law, Innovation and Technology, 2*, 203-221.

Kolber, A. J. (2007). Pain detection and the privacy of subjective experience. *American Journal of Law &*

Medicine, 33, 433-456.

Kozel, F. A., Johnson, K. A., Mu, Q., Grenesko, E. L., Laken, S. J., & George, M. S. (2005). Detecting deception using functional Magnetic Resonance Imaging. *Biological Psychiatry, 58*, 605-613.

Langleben, D. D., Schroeder, L., Maldjian, J. A., Gur, R. C., McDonald, S., Ragland, J. D., O'Brien, C. P., & Childress, A. R. (2002). Brain Activity during Simulated Deception: An Event-Related Functional Magnetic Resonance Study. *NeuroImage, 15*, 727-732.

McCabe, D. P., & Castel, A. D. (2007). Seeing is believing: the effect of brain images on judgments of scientific reasoning. *Cognition, 107*, 343-352.

Mobbs, D., Lau, H.C., Jones, O.D., & Frith, C.D. (2007). Law, Responsibility and the Brain. *PLoS Biology, 5(4),* e103.

Morell, L. C. (1998). New technology: experimental research on the influence of computer-animated display on jurors. *Southwestern University Law Review, 28*, 411-415.

Morse, S. (2006). Brain overclaim syndrome and criminal responsibility: a diagnostic note. *Ohio State Journal of Criminal Law, 3 (2)*, 397-412.

Nisbett, R., & Ross, L. (1980). *Human inference: strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Ochsner, K. N., Ludlow, D. H., Knierim, K., Hanelin, J., Ramachandran, T., Glover, G. C., & Mackey, S. C. (2006). Neural correlates of individual differences in pain-related fear and anxiety. *Pain, 120*, 69-77.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, and Winston.

Peyron, R., Laurent, B., & Garcia-Larrea, L. (2000). Functional imaging of brain responses to pain: A review and meta-analysis. *Journa of Clinical Neurophysiology, 30 (5)*, 263-288.

Reeves, D., Mills, M. J., Billick, S. B., & Brodie, J. D. (2003). Limitations of brain imaging in forensic psychiatry. *Journal of the American Academy of Psychiatry and the Law, 31 (1)*, 89-96.

Reyes, R. M., Thompson, W. C., & Bower, G. H. (1980). Judgmental biases resulting from differing availabilities of arguments. *Journal of Personality and Social Psychology, 39*, 2-12.

Roskies, A. L. (2007). Are neuroimages like photographs of the brain? *Philosophy of Science, 74*, 860-872.

Ross, L.D. (1977). The intuitive psychologist and his shortcomings: distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (173-220). New York: Academic Press.

Schweitzer, N.J. & Saks, M.J. (2007). The CSI Effect: Popular Fiction About Forensic Science Affects Public Expectations About Real Forensic Science. *Jurimetrics, 47*, 357-364.

Shelton, D. E., Kim, Y. S., & Barak, G. (2006). A study of juror expectations and demands concerning scientific evidence: does the "CSI effect" exist? *Vanderbilt Journal of Entertainment and Technology Law, 9 (2)*, 331-368.

Sinnott-Armstrong, W., Roskies, A., Brown, T., & Murphy, E. (2008). Brain images as legal evidence. *Episteme: A Journal of Social Epistemology, 5 (3)*, 359-373.

Spence, S. A., Farrow, T. F. D., Herford, A. E., Wilkinson, I. D., Zheng, Y. & Woodruff, P. W. R. (2001). Behavioural and functional anatomical correlates of deception in humans. *Neuroreport, 12 (13)*, 2849-2853.

Tovino, S. A. (2007). Functional neuroimaging and the law: trends and directions for future scholarship. *American Journal of Bioethics, 7 (9)*, 44-56.

Tyler, T. R. (2006). Viewing CSI and the threshold of guilt: managing truth and justice in reality and fiction. *The Yale Law Journal, 115*, 1050-1085.

Vedder, A., & Klaming, L. (2010). Human enhancement for the common good: Using neurotechnologies to improve eyewitness memory. *American Journal of Bioethics Neuroscience*, *1(3)*, 22-33.

Vincent, N. (2008). Neuroimaging and responsibility assessments. *Neuroethics, 1 (3)*, 199-204.

Weisberg, D.S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20 (3)*, 470-477.

Wolpe, P. R., Foster, K. R., & Langleben, D. D. (2005). Emerging neurotechnologies for lie-detection: promises and perils. *The American Journal of Bioethics, 5 (2)*, 39-49.

Chapter 6
Cross-cultural variation and fMRI lie-detection

Tommaso Bruni
University of Milan
Department of Medicine, Surgery and Dentistry
✉ tommaso.bruni@ifom-ieo-campus.it

Abstract As decidedly underscored by a recent editorial in Nature Neuroscience (2010), many experiments in cognitive neuroscience have been carried out with a sample that is not representative of the general human population, as the subjects are usually university students in psychology. The underlying assumption of this practice is that the workings of the brain do not vary much even when subjects come from different cultural groups. Recent research by Henrich et al. (2010) shows that this assumption is unwarranted. On several basic features of perception and cognition, Western university students turn out to be outliers relative to the general human population, so that data based on them should be interpreted with caution. In particular, this situation seems to provide an argument for questioning the conformity of functional Magnetic Resonance Imaging (fMRI) lie-detection to Federal Rule of Evidence 702 and *Daubert*. Deception is a social phenomenon and it is related to mental functions, such as theory of mind, for which cross-cultural variability at the neural level has been detected. Furthermore, culture is a multi-dimensional variable whose effects are diverse. Thus, the use of fMRI lie-detection in legal contexts may hinder the ascertainment of truth if the experimental results are not shown to be conserved in different cultures. Cross-cultural variability in neural activation patterns is just a facet of the broader issue of external and ecological validity for neuroscientific experiments on the detection of deception; nonetheless, fMRI lie-detection is unlikely to meet the *Daubert* standards if cross-cultural variation is not controlled by appropriate experiments.

Keywords fMRI, lie-detection, culture, cross-culturality, Daubert standard

Introduction

In this paper I discuss functional Magnetic Resonance Imaging (fMRI) lie-detection and claim that this technique should be used in courts only if its experimental basis includes checks for cross-cultural variation.

The concept of 'culture' refers to features of human groups that typically vary according to geographic areas and which depend on social learning; it includes shared attitudes, practices, and beliefs, together with languages and religions.

Cross-cultural variation in human psychology is pervasive (Norenzayan & Heine, 2005; Nisbett & Masuda, 2003) but it is rarely addressed in the behavioural sciences (Sears, 1986; Henrich, Heine, & Norenzayan, 2010). Cross-cultural variability in psychology corresponds, in some cases at least, to cross-cultural neural variability (for a review about cross-cultural neural variation, see Han & Northoff, 2008).

Lying is a social activity. As society and culture are closely related, deception is unlikely to be free of

cultural variation on both the psychological and the neural level. Moreover, culture possesses several dimensions (Hofstede, 2001), which are notoriously difficult to measure, so that it is much more complex to take this source of variation into account than others, such as for instance a mono-dimensional factor like age. For these reasons, the neuroscientists who are developing fMRI lie-detection should be aware of the problem and include cross-cultural experiments into their experimental strategies, in order to check if the Blood Oxygen Level Dependent (BOLD) activations that correlate with lying are conserved in different cultures. If this is not done, the experiments about fMRI lie-detection run the risk of having both a reduced ecological validity, i.e. the experimental settings are too heterogeneous relative to the parts of the real world they want to model, and a low external validity, i.e. the experiments are based on an idiosyncratic sample which is not representative of the general population. In this case the results would tell little about what happens outside the lab. If the experiments do not possess a sufficient degree of external and ecological validity, they are unlikely to provide error rates that are applicable to the real world and to gain general acceptance in the scientific community. But if the real-life error rates and general acceptance are absent, fMRI lie-detection will probably not be accepted as a valid expert testimony either in the jurisdictions that follow the *Daubert* ruling (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 1993) or in those that adopt the older *Frye* (*Frye v. United States*, 1923) test. This is because both tests take general acceptance into account and because one of the *Daubert* standards is the *"known or potential error rate"* in real life applications.

The concept of 'culture'

First of all, some words are due on the way I avail myself of the concept of 'culture' which is central to my overall argument. According to my working definition, 'culture' refers to some properties of human groups that depend on social learning. Languages, religions, shared attitudes and beliefs, family structures and hierarchies are all parts of culture. Culture varies not only moving from one social group to another, but also from an individual to another in the same group.[52] Both the geographical variability and the individual variability have behavioural consequences. For instance, Chua, Boland and Nisbett (2005) have demonstrated that Americans and Chinese feature different saccades[53] patterns when they are shown a picture composed by a salient object and a background: Chinese tend to focus more on the background than

[52] See for instance Haidt & Graham (2007) about the different moral principles used by liberals and conservatives.

[53] Saccades are quick and simultaneous movements of both eyes in the same direction. Human beings are usually not aware of performing saccades.

Americans. As to individual variation, priming for individualism or collectivism[54] performed on bicultural individuals, such as Japanese-Americans, modulates both their ways of self-description (general, context-free descriptions vs. contextual descriptions) and the corresponding BOLD signals in areas related to self-representation (Chiao et al., 2009). One of the major problems in dealing with culture as a factor of behavioural and neural variation is that culture is not easy to measure. One framework that I find helpful is Hofstede's (2001) five dimensional model, which collocates every culture along these dimensions:

1. Individualism – collectivism;
2. Small – large power distance: It measures the difference in power between the most and the least powerful members of the group. If power distance is large, the leaders of the group are much more powerful than the subordinates. If power distance is small, the leaders of the group are almost on the same level as subordinates;
3. Short – long term orientation: to what degree a group considers the remote future when making decisions;
4. Weak – strong uncertainty avoidance: how much a group is willing to take up risks;
5. Masculinity – femininity: here Hofstede uses the Western stereotypes as metaphors, without any commitment about the actual psychology of men and women. Masculinity symbolises an assertive and competitive stance, whereas femininity indicates a caring and modest attitude.

According to this model, every society is characterised by a set of five values that describe its position along the dimensions, but any individual in the society might depart from the group's values. For instance, the United States (US) are considered as one of the most individualistic societies in the world (Henrich et al., 2010), but a single US citizen can endorse collectivist values for a variety of reasons, such as religious tenets or family education.

Lastly, it must be understood that ethnicity is not a synonym of culture, since immigrants retain their ethnicity for some generations (as long as they have children with other immigrants coming from the same ethnic group), whereas they rapidly lose their original cultural traits (Heine & Lehman, 2004). Individual and intra-national variation also prevents us from identifying culture with nationality, even though nationality has a great influence on culture.

[54] Individualists think that people are independent from each other and that they are characterized by a context-independent set of personality traits. Collectivists see persons as interconnected and describe them as embedded in specific social situations, which constitute a part of their personality.

The sampling bias in the behavioural sciences

This being said, I can continue with the sampling bias. Most experiments in the 'behavioural sciences' (cognitive science, economics and psychology) are carried out on culturally homogeneous samples. Arnett (2008) has surveyed the articles of the main peer-reviewed journals in psychology in the 2003-2007 period and has found that 68% of the subjects come from the US. Furthermore, 67% of this US population is composed of university students who take psychology courses. Therefore, the bulk of experimental subjects in the behavioural sciences is composed by a very specific human group: undergrads in psychology.

On the one hand, this is an advantage, because very homogeneous samples allow the attribution of differences in the subjects' behavioural responses to the differences in the experimental conditions (e.g. distinct stimuli), which are manipulated by the researchers. Moreover, university students are easily available, cheap and permit a fast replication of the experiments.

On the other hand, this poses serious questions of generalisability of the experimental findings. How can a researcher be sure that the experimental results are valid under different cultural conditions? This risk is particularly serious if we take into account that university students are a very specific sample relative not only to the global human population, but also to the US population. As Rozin (in Henrich et al., 2010) has pointed out, the university student experiences a unique life transition from family life to a peer-centered life. Moreover, they usually earn little or no income, live in a very liberal, educated, and open-minded environment (the campus), and have not built their own family yet. Therefore, their behaviour on several accounts, such as economic decisions, is likely to be different even from that of the average US 30-year old person. This is evidenced by cross-cultural studies (Henrich et al., 2005) which show that the behaviour of university students coming from Western, industrialised countries on some economic games like the ultimatum game and the dictator game is very different from the behaviour found in many small-scale societies around the globe. A further consideration is that cultural variability does not only involve social behaviours like theory of mind[55] and its neural correlates (Kobayashi Frank & Temple, 2009) or economic behaviour, but has a much broader scope. For instance, on the behavioural level culture influences general strategies of reasoning (Nisbett, Nisbett, Peng, Choi, & Norenzayan, 2001), the performance on the visual 'rod-and-frame' task (Kitayama, Duffy, Kawamura, & Larsen, 2003), and the effectiveness of visual illusions (Segall, Campbell, & Herskovits, 1963).

Since one may understand the aim of the behavioural sciences as describing universal features of human behaviour and accounting for those features by means of appropriate theories, experiments that are

[55] Theory of Mind (ToM), or mentalising, is the ability to attribute mental states (both cognitive and affective) to other human beings.

carried out on a very specific sample are of little utility to the pursuit of such a purpose, at least as long as they are not repeated in different human groups that diverge culturally. It should be noted that universality must not be intended as a digital variable: there are discrete degrees of universality that can empirically be tested. For instance, a cognitive phenomenon can be present in almost all human groups, but perform different functions in different contexts, or it can be consistently present *and* robustly perform the same function in all contexts. Universality can be tested by means of three kinds of experiments: the two-cultures experiment, the triangulation study, and the cross-cultural survey (Norenzayan & Heine, 2005).

In a two-cultures experiment a determined response to an experimental setting is taken into account. Two cultures that differ on many cultural dimensions are examined and the experimenters check whether the effect is conserved. If it is, the experiment provides some evidence for some degree of universality; if it is not, the difference in the behavioural effects of the setting must be traced back to a cultural dimension. But since the two cultures that have been examined differ on many dimensions, identifying the dimension that is responsible for the variation is not straightforward. In order to do so, a triangular study is needed. Such a study must start from a theory that allegedly explains the previously tested effect and that allows researchers to make hypotheses as to which cultural dimension is responsible for the variation. Then the experimenters take into account three cultures that differ from each other along two theoretically relevant cultural dimensions. For instance, if the theory leads to the prediction that dimensions D_1 and D_2 may be relevant, the cultures will be selected in such a way that cultures C_1 and C_2 differ on D_1, and C_1 and C_3 differ on D_2. If the difference is spotted between C_1 and C_2, D_1 will be the relevant dimension; if the difference is found between C_1 and C_3, D_2 will be chosen as explanatory instead. Of course, it must be assured that in the different cultures, the experimental conditions are interpreted by the subjects in the same way and that the experimental protocol does not change.

A cross-cultural survey entails examining many human groups around the world, both in small-scale societies and in urban societies. If no differences are detected, it provides a strong evidence for some degree of universality. Nonetheless, it is costly and difficult to carry out, as experimental rigor cannot be maintained without considerable efforts when different research teams have to work in diverse environments. These cross-cultural investigations can be carried out by means of meta-analyses too, if sufficient data have already been gathered.

Furthermore, there are some types of behavioural research in which universality is not an issue, so that idiosyncratic samples can be used without any problems in these cases. As Gächter (in Henrich et al., 2010) correctly points out, US freshmen and sophomores can be very useful to falsify theories in behavioural economics. Falsification is about the research of counterexamples, not about generalisability, so that using undergrads as participants in an experimental study is appropriate when a study aims at falsification. Furthermore, students are bright enough to participate in these kinds of studies.

This being said about the sampling bias and how to address it, let us look at the part of the behavioural sciences that concerns me most in this paper: cognitive neuroscience. Here, the situation is probably even worse than in experimental psychology. According to Chiao (2009), 90% of the peer-reviewed neuroimaging studies come from Western industrialised countries. But the sampling bias would be a problem

only if significant evidence for cultural variability at the neural level has been gathered. Cultural neuroscience provides substantial evidence to this effect. I briefly review part of this evidence (for a more comprehensive review, see Han & Northoff, 2008). Gutchess, Welsh, Boduroğlu and Park (2006) have used fMRI to identify the neural correlates of a cross-cultural difference between Caucasian Americans and East Asians in image processing: Americans fixate a salient object more than East Asians. This proves that culture modifies neural function when non-verbal stimuli are processed.

Zhu and colleagues (2007) have found a differential activation of the Medial PreFrontal Cortex (MPFC), which explains the distinct construal of the self in American and Chinese subjects. In Americans, whose concept of self does not include intimate relatives, the MPFC is activated only in response to judgments concerning the subject himself, whereas in East Asian the same area of the brain also responds to stimuli concerning close relatives, such as the subject's mother.

Hedden and colleagues (2008) uncovered the neural correlates of another cross-cultural bias: East Asians are better than Americans at performing tasks that have contextual demands. Conversely, Americans are better than East Asians at ignoring the context if this is required.

By means of an fMRI study on Japanese bilinguals, Kobayashi, Glover and Temple (2006) have found differences in BOLD activation in Japanese and American cultures when subjects perform false belief tasks. False belief tasks are one of the main tests for theory of mind.

Wong and colleagues (2004) have shown in a Positron Emission Tomography (PET) study that the processing of auditory pitch patterns engages the left or right insular cortex when the pitch has a linguistic function, as in Chinese, or as not in English, respectively. This demonstrates that linguistic variation across cultures correlates with distinct neural correlates.

One can conclude that cross-cultural variation at the neural level concerns both basic brain functions, such as visual processing, and 'higher' functions such as self-construal. How can this problem be tackled? MRI scanners are expensive and it is difficult to find them in developing countries or to bring them to the homelands of small-scale societies. Conducting cross-cultural experiments in cognitive neuroimaging is therefore difficult. Nevertheless, East Asia provides a rich and industrialised area in which cultural variability relative to the West is still sufficiently high to make two-cultures neuroimaging experiments meaningful. One possible agenda for cultural neuroscience is to look for the neural correlates of the behavioural variation that has been found between East Asia and the US in cultural psychology.

The precise mechanisms by which culture can sculpt the human brain have not been elucidated yet, but the existence of brain plasticity is now an established fact. It has been studied both in the context of functional recovery after lesions (Wall, Xu, & Wang, 2002; Frost, Barbary, Friel, Plautz, & Nudo, 2003; Winship & Murphy, 2009) and in the context of learning (for instance Maguire et al., 2000). Brain plasticity yields a good theoretical framework to create detailed neural explanations of cross-cultural variability in behaviour, but cultural neuroscience still has a lot of work to do in order to reach the neurophysiological level on which small neural populations are taken into account. In addition, there are well-known and warranted ethical limitations to neurophysiological experimentation in humans.

In the next section I will examine fMRI lie-detection.

fMRI lie detection

Deception has been defined as *"a social behavior in which an individual attempts to persuade another to accept as true what the deceiver believes to be untrue"* (Ganis & Keenan, 2009, p. 465). Deception is a natural phenomenon that spontaneously develops in human beings (Spence et al., 2004). Deception as a mental task requires the suppression of a prepotent response, namely telling the truth; moreover, a new cognitive item, i.e. the lie, must be built up starting from the beliefs of the person to be deceived. Then the reactions of the deceived must be constantly monitored, so that consistency between the lie and their beliefs can be maintained. The lie can be very simple, as in cases in which one answers 'no' instead of 'yes' to a question, or quite complex when a whole piece of narrative must be devised to disguise the truth (Ganis, Kosslyn, Stose, Thompson, & Yurgelun-Todd, 2003). All this requires response inhibition, working memory, theory of mind, and other mental functions. Briefly, many high-order capacities of the human brain are engaged when one lies. Moreover, voluntary deception is essentially social: It is a way in which an individual manipulates her relationships with other human beings. Furthermore, there are many kinds of deception. In addition to the aforementioned distinction between structurally simple and complex lies, there are also the following differences:

1. self-related lies vs. other-related lies;
2. lies in which the subject says she did perform an action she has not carried out vs. lies in which the subject says she did not perform an action she has actually carried out (suggested by Kozel et al., 2009);
3. verbally expressed lies vs. non-verbally expressed lies;
4. well-rehearsed lies vs. improvised lies;[56]
5. lies in which a lot is at stake, in terms of rewards and risks, vs lies in which little is at stake.

Since the phenomenon is inherently social and there are many kinds of lies, the existence of a simple biological marker for all kinds of deception is unlikely (Sip, Roepstorff, McGregor, & Frith, 2008).

fMRI lie-detection has been tested in the lab using a variety of paradigms: a version of the polygraph Guilty Knowledge Test featuring playing cards (Langleben et al., 2002), a mock crime scenario (Kozel et al., 2005, 2009), autobiographical memories (Lee et al., 2002), and others. Nonetheless, all of these methods have common shortcomings.

Firstly, even if in some cases additional monetary rewards are promised to the subject if her lies are not detected by the experimenters (e.g. Kozel et al., 2005), the motivation to lie in the lab is very low relative

[56] For this last dichotomy Ganis & Keenan (2009) found different BOLD activations relative to a baseline constituted by telling the truth.

to some real life circumstances, in which the risks and gains of deceiving can be tremendous.

Secondly, in the lab, subjects are instructed to perform deception, so that the intention of lying is not spontaneous. Sip et al. (2008) claim that the lack of a voluntary intention to deceive prevents these experiments from studying deception. Instead, these experiments study *"some of the complex executive functions that are associated with the phenomenon [i.e. deception]"* (Sip et al., 2008, p. 48). This major shortcoming might be avoided by adopting an experimental setting in which subjects are put into a situation that indirectly induces them to be mendacious, on the lines of the experimental paradigm used by Greene and Paxton (2009). Nonetheless, to my knowledge no such study on deception has yet been conducted.

Thirdly, in all of these experiments the presence of lies is guaranteed by the experimental design, whereas in a real life setting the relevant issue is whether someone is lying or not. The findings of an experiment in which the presence of lies is secured cannot be extended to situations in which lies may or may not be there (Langleben & Dattilio, 2008).

Fourthly, the time between the fact that the subjects are questioned about and the scanning is usually short in lab settings (minutes or hours), whereas in real life it can be very long (months or years) (Spence et al., 2004).

Fifthly, the current paradigms compare two mutually exclusive conditions: Telling the truth vs. lying. In real life lies can be more nuanced: an account in which deception and what has actually happened are merged can be given (Spence et al., 2004). All of these factors contribute to creating a problem of external and ecological validity of fMRI lie-detection studies.

These experiments have identified a series of brain regions that correlate with lying in the experimental setting, i.e. that show an increased BOLD signal on a Lie minus Truth (henceforth written as Lie>Truth) contrast.[57] Many distinct brain regions have been indicated and researchers do not agree on which the most relevant regions are (Spence, 2008; Kozel et al., 2009), but some consistencies have been found (Monteleone et al., 2009). Firstly, there is no activation for Truth>Lie, showing that telling the truth is a baseline response. The regions that are regularly activated in Lie>Truth are associated with the cognitive functions that have been predicted to be involved in lying: response inhibition, working memory, theory of mind and others. The main areas that have been implicated are the MPFC (especially the ventromedial part known to be related to emotion processing, see Damasio, 1994), the orbitofrontal cortex, the anterior cingulate cortex, and the temporal parietal junction (TPJ). The dorsolateral prefrontal cortex seems to show an increased activation when lies are structurally complex, so that it may be related to the creation of a new

[57] fMRI investigation is normally based on the subtraction of the BOLD signal in the task condition from the BOLD signal in the control condition. In this case lying is the task and telling the truth is the control.

cognitive item (Spence et al., 2004). If only one area is examined, the best detection in terms of sensitivity at $p < 0.01$ is achieved by using the MPFC as in this way an accuracy of 71% is reached (Ganis & Keenan, 2009). Increasing accuracy above this level entails a rise in the number of false positives. If the number of liars in a population is very low, the number of false positives can be higher than the amount of the true positives, making the test useless. Therefore, false positives must be minimised if the technique is to be used in real life. The relatively low specificity of the test is due to the scarce specificity of the correlation between brain regions and deception. Those regions carry out many other functions and therefore their activation does not necessarily indicate deception. At the state of the art, fMRI lie-detection is only slightly more accurate than the traditional polygraphy (Simpson, 2008). Nonetheless, fMRI presents two advantages in comparison to polygraphy. Firstly, it measures a Central Nervous System (CNS) signal and not a Peripheral Nervous System (PNS) signal, therefore a closer correlate of behaviour and secondly, it is independent from arousal. In the next section, I will discuss some issues that arise from the potential application of this technique in criminal and civil proceedings.

fMRI lie-detection in judicial settings

I exclusively deal here with the US legal system. Firstly, I examine the legal standards that regulate the acceptance of scientific evidence and some recent decisions. Secondly, I argue that both external and ecological validity are central when the admission of fMRI lie-detection in court is discussed.

The admission of scientific evidence in US federal courts is regulated by Rule of Evidence 702, which concerns the testimony of scientific or technical experts. For the admission of the witness three conditions must be satisfied:

1. The testimony is based upon sufficient facts or data;
2. The testimony is the product of reliable principles and methods;
3. The witness has applied the principles and methods reliably to the facts of the case (Rule of Evidence 702).

These conditions are applied together with other standards that were fixed by two decisions of the Supreme Court of the US, the aforementioned *Frye v. United States* (1923) and *Daubert v. Merrell Dow Pharmaceuticals Inc.* (1993). The *Daubert* standards are valid in federal courts and in most state jurisdictions in the US. The *Frye* standard applies to the remaining state jurisdictions (among which are California and New York).

The *Frye* standard simply states that expert witnesses can be admitted in courts if *"the thing from which the deduction is made"* has *"gained general acceptance in the particular field in which it belongs"*.

Therefore, general acceptance on the part of the relevant scientific community is required. This general acceptance has not been reached in the case of fMRI lie-detection, as a 2008 editorial on Nature Neuroscience demonstrates. Moreover, a recent New York State decision in a civil case rejected the admission of fMRI lie-detection under *Frye*[58].

The *Daubert* standard attributes to the judge the role of gatekeeper with regard to expert witnesses. The *Daubert* standard must ensure that the testimony is relevant to the case and has been obtained by means of reliable methods. To ascertain this, a test with five non-exclusive and flexible prongs is proposed. The points are the following:

1. Empirical testing: the grounding theory must be falsifiable through experimentation;
2. Peer-reviewed publication of the scientific bases of the testimony;
3. Potential or known error rate of the procedure in real cases;
4. Existence of technical standards for the procedure;
5. General acceptance in the scientific community.

The admission of fMRI lie-detection under *Daubert* has been recently denied in the federal criminal case *USA v. Semrau*. The decision of Magistrate Judge Tu M. Pham[59] excludes fMRI lie-detection on two grounds:

1. Under Rule of Evidence 702 and *Daubert*, because *"there are no known error rates for fMRI-based lie detection outside the laboratory setting"*, because *"standards controlling the real-life application have not yet been established"*, because Dr. S. J. Laken, who performed the scans, violated his own protocols when he rescanned Dr. Semrau on a positive result for deception, and because *"fMRI-based lie-detection has not yet been accepted by the scientific community"*;
2. Under Federal Rule of Evidence 403, because the evidence was more prejudicial than probative, since the scans were taken by Dr. Laken without notifying the government. In this way, Dr. Semrau risked nothing in undergoing the tests, because positive results for deception would not have been released.

As Magistrate Judge Pham himself notices, the rejection of expert testimony is the exception, rather than the rule. This makes his decision particularly relevant.

---

58 Wilson v. Corestaff Services, decided May 14th 2010, Justice Robert J. Miller, 32996/07. Available at http://blogs.law.stanford.edu/lawandbiosciences/files/2010/06/CorestaffOpin1.pdf (accessed December 25th 2010). I thank Prof. Henry T. Greely for having pointed this case out to me.

59 USA v. Semrau, May 31st 2010, Magistrate Judge Tu M. Pham, No. 07-10074 MI/P. Available at http://blogs.law.stanford.edu/lawandbiosciences/files/2010/06/fMRI-Report-and-Recommendation1.pdf (accessed December 26th 2010).

The second point is not against fMRI lie-detection per se, but it concerns the contingent circumstances of the *USA v. Semrau* scans, so that it is not relevant for our discourse. The first point in contrast is paramount: fMRI lie-detection is not admitted *inter alia* because there are no reliable error rates. This is due to the fact that most of the current experimental work tells us little about real life application of this technique. Concerns about external and ecological validity are particularly relevant in legal contexts. In fact, judicial applications of fMRI lie-detection might be conducted on people who are medicated, who may have a psychiatric history, who are unwilling to cooperate, and who may try to use countermeasures to the test. However, the technique has been tested so far on subjects without any psychiatric condition, present or antecedent, who are unmedicated, and who are willing to follow the instructions of the experimenters.

Another factor, which is specific to legal settings, must be taken into account. As Simpson (2008) correctly points out, the current paradigm of fMRI lie-detection focuses on functions such as response inhibition and correlated regions in the brain. But response inhibition is likely to be very often engaged by a defendant in a criminal trial, since defendants must be circumspect about what they say in courts and repress feelings of outrage at accusations, if they are not guilty. Therefore, if lie-detection is carried out in the context of a criminal trial, response inhibition seems to be an unreliable marker for deception. This shows again that the experimental paradigms used so far might have little bearing on how deception outside the lab, and specifically in a court, is detected.

This allows us to conclude that fMRI lie-detection is unlikely to be accepted under the Federal Rule of Evidence 702 and *Daubert* unless problems relative to external and ecological validity are solved.

In addition to this, I argue that cross-cultural validity is an important issue in this set of problems and that it needs to be addressed if fMRI lie-detection is to enter courts under the laws currently in force. For this claim I present four arguments. First, as briefly mentioned above, Kobayashi and her co-workers (2006, 2007) have shown that different areas of the brain are activated in East Asians and Caucasian Americans when they perform the false belief task. In particular, the TPJ activation would be culture-dependent and specific to English-speaking cultures (Kobayashi Frank & Temple, 2009). Nonetheless, Adams et al. (2009) have found a consistent activation of the posterior part of the superior temporal sulcus (pSTS), which partially overlaps with the TPJ, in both American and Japanese subjects. But Adams and co-workers (2009) used a different task than the one that Kobayashi Frank and Temple (2009) availed themselves of. The task used by Adams et al. (2009) is non-verbal and based on eye stimuli, whereas the false belief task is normally verbal. The different results of the two groups might be due to the distinct stimuli that were used. Despite this, researchers agree that the brain areas activated during ToM tasks depend on cultural background. Furthermore, culture impacts on ToM in other ways. For instance, people are better at detecting mental states in targets belonging to their in-group relative to out-group members and different areas of the MPFC are activated when subjects are asked to use ToM on targets that are respectively similar or dissimilar to themselves (Mitchell, Macrae, & Banaji, 2006). From this I can conclude that the neural underpinnings of ToM show a high degree of cross-cultural variation. If ToM is a necessary cognitive component of deliberate deception (and it is difficult to think it is not, as deception requires a manipulation of another's beliefs), this cultural variation is likely to be shared by deliberate deception too, even though experiments are needed to

confirm this theoretical prediction. And if BOLD patterns varied significantly across cultures in deception, fMRI lie-detection would risk being unreliable when a single experimental paradigm is used on subjects belonging to different cultures.

Secondly, the sheer number of immigrants in the US constitutes a good argument for cross-cultural checks. There were 38 million first-generation immigrants in the US in 2007 (Segal, Elliott, & Mayadas, 2010), amounting to about 12% of the US population. If fMRI lie-detection was used in court, more than one out of ten suspects could potentially show cultural variability in the neural correlates of lying, assuming *arguendo* that immigrants end up under trial or in civil litigation with the same frequency as the general US population. Therefore, if cross-cultural validity of fMRI lie-detection is not checked by means of appropriate experiments, errors could be widespread, leading to sub-optimal outcomes of judiciary procedures. Of course, the real amount of cultural variation in the neural correlates of deliberate deception cannot be estimated without actual cross-cultural experiments.

Thirdly, culture is different from other forms of variation in that it has many dimensions and components, together with a degree of individual variation. Unlike age, which is mono-dimensional, culture is manifold and therefore difficult to handle. Each cultural dimension could have a different effect on the neural correlates of voluntary deception, so that adapting the experimental setting of the technique to the culture of the individual to be tested might prove a daunting task. A careful measurement of the different cultural dimensions of the individual might be required. If the BOLD signals found during deception varied with culture, it could be extremely complex to devise an fMRI lie-detection technique suitable to use in courts under *Daubert*, since the error rate would be high. On the contrary, assuming *arguendo* that the neural correlates of deception vary with age, it might be easier to modify the experimental paradigm to factor this source of variability in, since the age of every person can be easily assessed.

Fourthly, the social nature of deception makes it theoretically likely that culture plays a big role in shaping this phenomenon, as culture, unlike for instance age, is a source of variation that results from human sociality. Deception requires a continuous surveillance on the beliefs of the deceived, an estimation of the long-term consequences of deception, and a maintenance of trust by means of pseudo-cooperation (Sip et al., 2008). Variations in belief systems and in what is considered to be advantageous or disadvantageous could thus make substantial changes in the psychological nature of deception. Since psychological differences are often coupled with underlying neural differences, this variability would affect BOLD signalling as well.

I am not claiming that the neural correlates of deception vary with culture, but that from the theoretical point of view this is likely to be the case. This hypothesis must be addressed by means of cross-cultural experiments.

There are of course many other legal and ethical issues that are raised by fMRI lie-detection in a legal setting. Firstly, there is the concern that lie-detection might illegitimately reduce the prerogatives of the finder of fact, who has *inter alia* the role of assessing the credibility of witnesses. Then, we find the so-called 'CSI effect', as Simpson (2008) states it:

> The aura of big science and high technology surrounding complex and expensive tests may lead to an overestimation of the reliability and utility of fMRI lie detection among lay people, including law enforcement personnel and other investigators, judges, and jurors (Simpson, 2008, p. 496).

A third issue is related to the Fourth and Fifth Amendment of the US Constitution. Concerning the Fourth Amendment, fMRI lie-detection could be seen as an unreasonable search and seizure and as a violation of the individual's cognitive freedom (Wolpe, Foster, & Langleben, 2005) if it is performed without an appropriate warrant. Concerning the Fifth Amendment, forcing the defendant to undergo lie-detection might be interpreted as an instance of self-incrimination. These problems are very important and must be carefully considered when discussing the ethical and legal acceptability of fMRI lie-detection in court. Nonetheless, dealing with them in depth would lead me astray as they are not connected with cross-culturality and because they play a minor role in the *USA v. Semrau* landmark decision. In the next section I will address some possible objections that can be made against my arguments.

Discussion of objections

Kozel and colleagues (2005, 2009) have done much to tackle the external and ecological validity problem, of which the cross-culturality issue constitutes a part. In particular, Kozel et al. (2005, 2009) used quite diverse samples, which cover a broad age interval, different ethnicities, professions and levels of education. None of these factors is significantly correlated with the results of the experiments. This strengthens the external validity of the study. Moreover, the more recent study makes use of a mock sabotage scenario which is much closer to a real life situation than the previous scenarios (subjects are asked to go to a separate building, find a CD containing evidence of a crime, devise a way to destroy it, and go back to the experimenter; a phone rings in the room where the CD is kept in order to enhance emotional stress). This increases ecological validity. Finally, Kozel et al. (2009) addressed the aforementioned problem of the time interval between the relevant action (in this case the sabotage) and the scanning. They have brought the time-lapse to 105 hours, but it is not clear whether this time interval is sufficient to solve the issue, as in real legal applications the time would probably be much longer. Nevertheless, this is another step towards a greater ecological validity.

Does the work of Kozel and colleagues (2005, 2009) undermine the legitimacy of requesting cross-cultural checks? I argue it does not, because experiments were carried out in a US university, using a sample of general US residents, and because ethnicity cannot be identified with culture. For instance, African Americans and Western Africans may be very similar from the ethnic point of view, but they undoubtedly differ a lot along many cultural dimensions. Even though the populations used by Kozel et al. (2005, 2009) are diverse, they are likely to be relatively homogeneous from the cultural point of view, as they are mostly composed of people born and raised in the US. If a significant proportion of first-generation immigrants had been included, more precise conclusions about the need of cross-cultural checks could have been drawn. This does not detract from the value of the studies conducted by Kozel and colleagues (2005, 2009), which according to my view is the only research on fMRI lie-detection that takes the important problem of external

and ecological validity into account.

A second important objection refers to the current practices of lie-detection and the role cross-culturality plays in them. Juries currently evaluate the truthfulness of witnesses not only by the plausibility of their statements and by the consistency of their account, but also through a gamut of behavioural cues (fidgeting, speed of speech, keeping eye contact with the jury, and so on) whose reliability is not above chance (Rand, 2000; Ganis & Keenan, 2009; Schauer, 2009). Nevertheless, jurors may consider these cues to be quite reliable. It is likely that these clues undergo cultural variation[60]: What is considered to be a sign of reliability can obviously change across cultures. Specific evidence to this effect is available: e.g. Bond and colleagues (1990) have shown that American and Jordanian observers rely on partially different sets of cues to detect deception. Therefore, it may be the case that for instance immigrant defendants are already disadvantaged in trials because they do not know what kind of demeanour they are supposed to keep in front of the jury in order to look truthful. They might abide the unwritten rules of their home culture and use a body language that does not match the expectations of the jurors. As Rand (2000) points out, truthful African-American witnesses could be seen as liars by Caucasian American jurors because of this 'Demeanor Gap'. Therefore, so the objection goes[61], we already have a cross-culturality problem in lie-detection. This renders a cross-culturality problem in fMRI less important, as we would simply not solve a problem we already have in the current situation. Continuing on the same lines, as fMRI is much more accurate than behavioural cues as a tool of lie-detection, it would be a good idea to adopt it, since it simply keeps the cross-culturality issue unsolved, but provides a much higher detection rate. According to Bold (1990), both American and Jordanian observers have detected lies with an accuracy rate of slightly more than chance (about 54%). FMRI lie-detection reaches more or less 70% (Ganis & Keenan, 2009) without false positives.

To this argument I respond that the CSI effect' generates a big difference between lie-detection by bodily cues and fMRI lie-detection. The jurors would consider the latter as 100% accurate. Jurors would probably have some doubts about truthfulness assessment via body language and voice pitch, whereas fMRI lie-detection seems to eliminate all uncertainty. Therefore, cultural biases in current trials produce milder harms than those that would result from alleged cultural biases in fMRI lie-detection. Given the CSI effect, false positives in fMRI lie-detection might cause severe trouble. Then, if we move from this level to the legal standards for admission, we notice that this argument is irrelevant for the conformity of fMRI lie-detection to the *Daubert* requirements. Already having a problem in our current practices does not make the case for fMRI lie-detection relative to *Daubert* easier.

---

60 I thank one anonymous referee for pointing this out to me.

61 It should be noted that this objection is not proposed by Rand (2000), but is rather a theoretical reconstruction of a possible line of argument.

A third objection claims that we should not wait to use fMRI lie-detection, since:

1. The current practices of lie-detection are really bad (see the behavioural cues above);
2. There is a huge societal demand of lie-detection (Langleben, 2008), such that the US government continues to use polygraphy even though its accuracy is far from being perfect;
3. Because current methods of extracting information, such as waterboarding, are cruel and violate human rights (Spence et al., 2004);
4. Cross-cultural checks would take a long time, a time that we cannot allegedly afford to lose.

My response to this is threefold. Firstly, lie-detection is not mind-reading, which is at present a totally futuristic technology, so that fMRI lie-detection cannot be considered an information extraction technique. Therefore it is improbable that fMRI will replace forms of torture in the near future. Secondly, using fMRI lie-detection without checking for generalisability might entail sub-optimal outcomes of the judicial processes, such as punishment of the non-guilty, acquittal of the guilty, and payment of undue compensations. It is not clear if an early use of fMRI lie-detection would make criminal or civil trials better in the present situation. Given the possibility of a 'CSI effect' with regard to fMRI lie-detection, the risk that it would not improve trials is significant. Thirdly, the societal demand for lie-detection can be questioned from two perspectives. On the one hand, it might be argued that the demand for lie-detection is an American peculiarity, maybe even an obsession, as historian Ken Alder (2007) has claimed. On the other hand, this demand can be cast into doubt from the ethical point of view. Is this demand warranted? What kind of balancing between security and individual liberties do we want to adopt? How are we to interpret the citizens' cognitive freedom? This is an issue I cannot discuss here, but of course the legitimacy of this societal demand cannot be taken for granted. As a matter of fact, neuroethics experts like Levy (2007) have argued that early adoption is the main risk when neuroscientific lie-detection is discussed.

The fourth objection comes from Schauer (2009): he denies the relevance of any kind of scientific considerations concerning external and ecological validity. Schauer argues:

> If the ease of telling an instructed lie in the laboratory correlates with the ease of telling a real lie outside the laboratory, research on instructed lies is no longer irrelevant to detecting real lies. With any positive correlation between instructed and real lies, experiments on the former will tell us something about the latter, and whether that 'something' is enough depends on the uses for which the research is employed. That which is inadequate for scientific publication or criminal prosecution might be sufficient for a defendant seeking to suggest reasonable doubt (Schauer, 2009, p. 102).

The overall point Schauer (2009) is making is that legal and not scientific standards matter in assessing evidence in courts. Both external and ecological validity are scientific standards and therefore are allegedly not relevant in a legal context. It is sufficient to have 'something' that binds the lab setting and real life to permit some use of the experimental results. Therefore, even though fMRI lie-detection has some problem of external validity on the scientific level, it could be used in legal settings, such as civil litigation, where the standards of evidence are not *"beyond a reasonable doubt"*, but *"a preponderance of evidence"* or *"a reasonable suspicion"*.

What surprises in this account is that Schauer's (2009) arguments ignore the Federal Rule of Evidence 702 and *Daubert* altogether. These norms state that the legal standards for expert witness *are* at least in part scientific standards. If these *scientific* standards are not met, the evidence cannot be *legally* admitted. This also applies to all arguments for differential application of fMRI lie-detection (admissible in civil cases but not in criminal cases; admissible for the defence but not for the prosecution inside criminal cases) at the federal level. The Federal Rule of Evidence governs proceedings in the federal courts of the US whatever the case at issue (civil or criminal), so that there seems to be no room for differential application. Judges, as gatekeepers, must decide on admissibility on a case per case base: every use must be separately evaluated in its specific context. Nevertheless, the federal judge must abide the Federal Rules of Evidence and *Daubert* in doing so. Therefore, external validity cannot be dismissed as being merely scientific and not legally relevant.

The fifth and last objection underlines that no cross-cultural variation was found by using behavioural tests for deception such as the polygraph[62] and that no cross-cultural neural variation has been reported so far for deception paradigms. The two points seem to show that the worry I am expressing I implausible. If there is no behavioural cross-cultural variation for earlier lie-detection techniques and neural cross-cultural variation has never been detected in deception, maybe there is no good reason to presume that the latter can be a problem for fMRI lie-detection. To this I reply that it is difficult to find cross-cultural variation in a test if this is not explicitly searched for, especially if there is no way to double-check for the correctness of the result. If a polygraph is used in a real-life setting and signals a suspect or witness as a liar, it is not so easy to check whether the machine is right or not, because the reliability of the subject was doubtful in the first place. Furthermore, behavioural researchers often start from the implicit assumption that cross-cultural variation is negligible, so that they do not notice this phenomenon unless it is macroscopic. To the best of my knowledge, neural cross-cultural variation in deception is yet to be tested. I would welcome any experimental attempt either to show its presence or to demonstrate its nonexistence.

Conclusion

The long and the short of this paper is that cross-cultural experiments on fMRI lie-detection should be performed before this technique enters courts, because the lab experiments with US citizens risk having an unacceptably low external validity. As a matter of fact, I suggest the technique cannot live up to the *Daubert* standards without such checks, because no error rate calculated in the lab can be projected onto real life without them. I do not take any position about the ethical acceptability of fMRI lie-detection, but argue that

62 To the best of my knowledge no cross-cultural variation is explicitly reported in the polygraph literature.

more neuroscientific research is needed (not only in the cross-cultural field) in order to assess its full potential both legally and morally. I therefore encourage and endorse more funding for fMRI lie-detection research. Only sound and carefully conducted empirical research can lead to new forensic technologies that can be useful to ascertain the truth and to justly determine legal proceedings.

References

Adams, R. B., Rule, N. O., Franklin, R. G., Wang, E., Stevenson, M. T., Yoshikawa, S., Nomura, M., Sato, W., Kveraga, K., & Ambady, N. (2009). Cross-cultural reading the mind in the eyes: An fMRI investigation. *Journal of Cognitive Neuroscience, 22(1),* 97-108.

Alder, K. (2007). *The lie detectors: The history of an American obsession*. New York: The Free Press.

Arnett, J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist, 63(7),* 602-14.

Bond, C. F., Omar, A., Mahmoud, A., & Bonser, R. N. (1990). Lie detection across cultures. *Journal of Nonverbal Behavior, 14(3),* 189-204.

Chiao, J. Y. (2009). Cultural neuroscience: a once and future discipline. *Progress in Brain Research, 178,* 287-304.

Chiao, J. Y., Harada, T., Komeda, H., Li, Z., Mano, Y., Saito, D., Parrish, T. B., Sadato, N., & Iidaka, T. (2009). Dynamic cultural influences on neural representations of the self. *Journal of Cognitive Neuroscience, 22(1),* 1-11.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *PNAS 102(35),* 12629-12633.

Damasio, A .(1994). *Descartes' Error*. London: Vintage.

Daubert v. Merrell Dow Pharmaceuticals. 516 U.S. 869 (1993).

Editorial (2008). Deceiving the law. *Nature Neuroscience, 11 (11),* 1231.

Editorial (2010). The university student as a model organism. *Nature Neuroscience, 13 (5),* 521.

Frost, S. B., Barbary, S., Friel, K. M., Plautz, E. J., & Nudo, R. J. (2003). Reorganization of remote cortical regions after ischemic brain injury: a potential substrate for stroke recovery. *Journal of Neurophysiology, 89,* 3205-3214.

Frye v. United States. 293 F. 1013 (D.C. Cir. 1923).

Ganis, G., Kosslyn, S. M., Stose, S., Thompson, W. L., & Yurgelun-Todd, D. A. (2003). Neural correlates of different types of deception: an fMRI investigation. *Cerebral Cortex, 13(8)*, 830-836.

Ganis, G., & Keenan, J. P. (2009). The cognitive neuroscience of deception. *Social Neuroscience, 4(6),* 465-472.

Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions, *PNAS, 106(30),* 12506-12511.

Gutchess, A. H., Welsh, R. C., Boduroğlu, A., & Park, D. C. (2006). Cultural differences in neural function associated with object processing. *Cognitive, Affective, and Behavioral Neuroscience, 6(2),* 102-109.

Haidt, J., & Graham, J. (2007). When morality opposes justice. Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research, 20(1),* 98-116.

Han, S., & Northoff, G. (2008). Culture-sensitive neural substrates of human cognition: a transcultural neuroimaging approach. *Nature Reviews Neuroscience, 9,* 646-654.

Hedden, T., Ketay, S., Aron, A., Markus, H. R., & Gabrieli, J. D. E. (2008). Cultural influences on neural substrates of attentional control. *Psychological Science, 19(1),* 12-17.

Heine, S. J., & Lehman, D. R. (2004). Move the body, change the self: Acculturative effects on the self-concept. In M. Schaller & C. Crandall (Eds), *Psychological foundations of culture* (305-331). Mahwah, NJ: Erlbaum.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., & Tracer, D. (2005). “Economic man” in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences, 28(6),* 795-855.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences, 33,* 61-135.

Hofstede, G. (2001). *Culture’s consequences: Comparing values, behaviors, institutions and organizations across nations*. Thousand Oaks, CA: Sage Publications.

Kitayama, S., Duffy, S., Kawamura, T., & Larsen, J. T. (2003). Perceiving an object and its context in different cultures: a different look at new look. *Psychological Science, 14(3),* 201-206.

Kobayashi, C., Glover, G. H., & Temple, E. (2006). Cultural and linguistic influence on neural bases of ‘‘theory of mind’’: An fMRI study with Japanese bilinguals. *Brain and Language, 98,* 210–220.

Kobayashi, C., Glover, G. H., & Temple, E. (2007). Cultural and linguistic effects on neural bases of ‘‘theory of mind’’ in American and Japanese children. *Brain Research, 1164,* 95-107.

Kobayashi Frank, C., & Temple, E. (2009). Cultural effects on the neural basis of theory of mind. *Progress in Brain Research, 178,* 213-223.

Kozel, F. A., Johnson, K. A., Mu, Q., Grenesko, E. L., Laken, S. J., & George, M. S. (2005). Detecting deception using functional magnetic resonance imaging. *Biology and Psychiatry, 58,* 605-613.

Kozel, F. A., Johnson, K. A., Grenesko, E. L., Laken, S. J., Kose, S., Lu, X., Pollina, D., Ryan, A., & George, M. S. (2009). Functional MRI detection of deception after committing a mock sabotage crime. *Journal of Forensic Science, 54(1),* 220-231.

Langleben, D. D. (2008). Detection of deception with fMRI: Are we there yet? *Legal and Criminological Psychology, 13(1),* 1-9.

Langleben, D. D., Schroeder, L., Maldjian, J. A., Gur, R. C., McDonald, S., Ragland, J. D., O’Brien, C. P., & Childress, A. R. (2002). Brain activity during simulated deception: An event-related functional magnetic resonance study. *NeuroImage, 15,* 727-732.

Langleben, D. D., & Dattilio, F. M. (2008). Commentary: The future of forensic functional brain imaging. *Journal of the American Academy of Psychiatry and The Law, 36(4),* 502-504.

Lee, T. M. C., Liu, H.-L., Tan, L.-H.,Chan, C. C., Mahankali, S.,Feng, C.-M., Hou, J., Fox, P. T., & Gao, J.-H. (2002). Lie detection by functional Magnetic Resonance Imaging. *Human Brain Mapping, 15*, 157-164.

Levy, N. (2007). *Neuroethics. Challenges for the 21st Century*. Cambridge: Cambridge University Press.

Maguire, E. A., Gadian, D. J., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *PNAS, 97(8),* 4398-4403.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50,* 655-663.

Monteleone, G. T., Phan, K. L., Nusbaum, H. C., Fitzgerald, D., Irick, J.-S., Fienberg, S. I., & Cacioppo, J. T. (2009). Detection of deception using fMRI: Better than chance, but well below perfection. *Social Neuroscience, 4(6),* 528-538.

Nisbett, R. E., & Masuda, T. (2003). Culture and point of view. *PNAS, 100 (19),* 11163–11170.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus

analytic cognition. *Psychological Review, 108(2),* 291-310.

Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin, 131(5),* 763-778.

Rand, J. W. (2000). The demeanor gap: Race, lie detection, and the jury. *Connecticut Law Review, 33(1),* 1-76.

Rule of Evidence 702. Retrieved http://www.law.cornell.edu/rules/fre/rules.htm (accessed December 28th 2010).

Schauer, F. (2009). Neuroscience, lie-detection, and the law. *Trends in Cognitive Sciences, 14(3),* 101-103.

Sears, D. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature, *Journal of Personality and Social Psychology, 51(3),* 515-30.

Segal, U. A., Elliott, D., & Mayadas, N. S. (2010). *Immigration worldwide: Policies, practices, and trends*. New York: Oxford University Press.

Segall, M. H., Campbell, D. T., & Herskovits, M. J. (1963). Cultural differences in the perception of geometric illusions. *Science, 139,* 769-771.

Simpson, J. R. (2008). Functional MRI lie-detection. Too good to be true? *Journal of the American Academy of Psychiatry and The Law, 36(4),* 491-498.

Sip, K. E., Roepstorff, A., McGregor, W., & Frith, C. D. (2008). Detecting deception: the scope and limits. *Trends in Cognitive Sciences, 12 (2),* 48-53.

Spence, S. A., Hunter, M. D., Farrow, T. F. D., Green, R. D., Leung, D. H., Hughes, C. J., & Ganesan, V. (2004). A cognitive neurobiological account of deception: Evidence from functional neuroimaging. *Philosophical Transactions of the Royal Society of London B, 359,* 1755-1762.

Spence, S. A. (2008). Playing Devil's advocate: The case against fMRI lie-detection. *Legal and Criminological Psychology, 13,* 11-25.

Wall, J. T., Xu, J., & Wang, X. (2002). Human brain plasticity: An emerging view of the multiple substrates and mechanisms that cause cortical changes and related sensory dysfunctions after injuries of sensory inputs from the body. *Brain Research Reviews, 39,* 181-215.

Winship, I. R., & Murphy, T. H. (2009). Remapping the somatosensory cortex after stroke: Insight from imaging the synapse to network. *Neuroscientist, 15(5),* 507-524.

Wolpe, P. R., Foster, K. R., & Langleben, D. D. (2005). Emerging neurotechnologies for lie-detection: Promises and perils. *American Journal of Bioethics, 5(2),* 39-49.

Wong, P. C. M., Parsons, L. M., Martinez, M., & Diehl, R. L. (2004). The role of the insular cortex in pitch pattern perception: The effect of linguistic contexts. *Journal of Neuroscience, 24(41),* 9153-9160.

Zhu, Y., Zhang, L., Fan, J., & Han, S. (2007). Neural basis of cultural influence on self-representation. *Neuroimage, 34,* 1310-1316.

Section C: Neuroscience and enhancement

Chapter 7
Moral enhancement: What is it and do we want it?

Anna Pacholczyk
University of Manchester
School of Law
✉ annapacholczyk@gmail.com

Abstract Building on the achievements of disease-oriented research, the coming decades will witness an explosion of biomedical enhancements to make people faster, stronger, smarter, less easily distracted and forgetful, happier, prettier, and live even longer. Recently, there has been a new arrival on the enhancement scene – moral and social enhancement. In one of the most significant works on moral enhancement to date, Julian Savulescu and Ingmar Persson (2008) suggest that *"the core moral dispositions...have biological basis and, thus, in principle should be within the reach of biomedical and genetic treatment"* (Savulescu & Persson, 2008, p. 172) although they question to what extent these interventions can be done in practice. I explore what we mean by moral enhancement and draw some distinctions that will help us avoid confusion when talking about the matter. Next, I suggest that the pessimistic view of the plausibility of moral enhancement stems from having much higher expectations about the effectiveness of morally modifying interventions. However, if we make our expectations comparable to those we have of cognitive enhancement or pharmacological treatment, then current research in the field of neuroscience of morality suggests that relatively efficient interventions are already here or will be possible in the near future. Next, I draw our attention to the plethora of potential targets of enhancement and discuss oxytocin as a potential moral enhancer. Finally, I highlight and explore possible problems with morally enhancing interventions, such as the threat to freedom and problems of application stemming from the lack of consensus about what is morally permissible and obligatory. I suggest that even if we accept that there are cases of fundamental moral disagreement, the problem may be much less serious then it first appears.

Keywords enhancement, pharmacological enhancement, moral enhancement, morality, ethics

Introduction

Moral enhancement is, at least *prima facie*, not susceptible to some of the objections to cognitive enhancement. Making people more moral seems to be beneficial to society, and thus the typical concerns raised in connection to cognitive enhancement – that enhancement is beneficial for the subject of enhancement but harmful to others – appears not to apply. Tom Douglas (2008) used the example of moral

enhancement to refute what he called the *bioconservative thesis*[63]. Douglas (2008) argues that the bioconservative thesis is false, given that there is at least one enhancement that would be morally permissible, namely enhancement of moral motives. In the second significant work that considered moral enhancement, Julian Savulescu and Ingmar Persson (2008) suggest that although *"the core moral dispositions...have a biological basis and, thus, in principle should be within the reach of biomedical and genetic treatment"* (Savulescu & Persson, 2008, p. 172), sufficiently effective interventions are not within our reach in the near future. The purpose of this paper is to extend the scope of the debate and address some of the objections raised as to plausibility and desirability of moral enhancement.

In the first part of this paper, I will examine what 'moral' in moral enhancement can mean, and suggest that there are three main ideas that moral enhancement can refer to: a morally permissible or even obligatory enhancement of any kind, an enhancement that would make people more moral in some way and a beneficial intervention that affects moral functioning. Secondly, I discuss the plausibility of moral enhancement. I look at Savulescu and Persson's (2008) paper on moral enhancement and examine what moral enhancement is expected to do. I suggest that the pessimistic view of the plausibility of moral enhancement expressed by these authors stems from having overly high expectations about the effectiveness of morally modifying interventions. However, if we make our expectations comparable to those we have of cognitive enhancement, then current research in the field of neuroscience of morality and pro-social behaviour suggests that relatively efficient interventions are already here or will be possible in the near future. I then discuss some of the possible ways in which moral functioning could be influenced and discuss whether moral enhancement is something we may want given the existence of moral disagreement and an alleged threat to freedom posed by moral enhancement.

What is moral enhancement?

What does the term 'moral' in 'moral enhancement' mean?

In this section I explore what we mean by moral enhancement. I suggest some distinctions that might help us avoid confusion when talking about the matter and propose that 'moral' can have three meanings in this context. As such, moral enhancement can be understood as an ethically desirable enhancement of any capacity, an ethically desirable enhancement of a moral sphere, or an enhancing intervention affecting the moral sphere that brings an overall benefit to the subject of enhancement.

[63] "Even if it were technically possible and legally permissible for people to engage in biomedical enhancement, it would not be morally permissible for them to do so" (Douglas, 2008, p. 228).

'Moral' enhancement as enhancement that is morally desirable

When we say 'moral enhancement' what we could mean is that it is an enhancement of any kind that is morally desirable. We may here think about enhancement that would result – other things being equal – in a better world. Vaccinations for smallpox resulted in the eradication of this disease (Eyler, 2003), and most would agree that the world without suffering and deaths brought by smallpox is better than an otherwise identical world with this disease. We often think that the increase in average life expectancy that took place in the past century was a good thing and it is, at the very least, an ethically permissible goal of the state to promote longevity, especially if it is accompanied by a good quality of life (Harris, 2007). Some have proposed that cognitive enhancement is not only permissible but that there can be a duty to enhance (Chan & Harris, 2007). Enhancements can therefore be said to be moral in the sense of being morally permissible or even morally obligatory. Thus, 'moral' in the first sense refers solely to our ethical appraisal of a given enhancement. However, when we say 'moral enhancement' we might mean something very different.

Moral enhancement as a change in some aspect of morality that results in a morally better person

Moral enhancement can also refer to making people more ethical, thus making them morally better in some sense. This is what Savulescu and Persson (2008) had in mind when they proposed that moral enhancement could, theoretically, be an answer to an alleged increased risk posed by the cognitively enhanced and morally corrupt minority.

Being moral is a complex ability and there is a wide range of potentially enhancing interventions. Thus, making morally better people could include making people more likely to act on their moral beliefs, improving their reflective and reasoning abilities as applied to moral issues, increasing their ability to be compassionate and so on. We could also focus on a number of aspects of being moral - acting in a moral way, being more virtuous, having better moral motivations and so on. Some of those possibilities will be explored in section 2.

There are two parts to the idea of moral enhancement understood as making people morally better: the factual claim that the enhancement in question in some way affects the moral sphere, and a normative claim about whether that intervention makes for a morally better person. Those two components may at first seem necessarily coexisting. But the distinction between the factual and a normative claim about moral enhancement is important, as our discussion may be constructed differently depending on whether we take the factual and normative or only factual claim as a basis for our discussion. The first reason for that is pragmatic – making sure that we keep this distinction in mind can make the discussion clearer. Secondly, considerations of moral enhancement based solely on the factual claim are interesting in their own right, and we would be missing an important part of the enquiry if we focused only on enhancement understood as making people more moral.

Moral enhancement as a beneficial change in the sphere of morality

As mentioned before, 'moral' in the phrase 'moral enhancement' can have a descriptive function and refer to, for example, a certain aspect of our cognition. A cognitive approach to moral enhancement would

therefore be based on the assessment of cognitive functions and regions implicated in moral reasoning, decision-making, acting and so forth, and on the assessment of how these functions can be modulated. On that view, whether an intervention is a moral enhancement depends on whether it affects relevant cognitive processes and behaviours. We may construct the moral sphere narrowly or widely but in this paper I will not consider this issue in great depth. Secondly, it depends on whether the modification of function counts as an enhancement. In the next section I propose and discuss the understanding of enhancement as improvement.

Enhancement as improvement

Elsewhere John Harris and I have proposed understanding enhancement as an improvement brought about by a change in a characteristic or function and an intervention that is overall beneficial (Pacholczyk & Harris, forthcoming). Does understanding enhancement as improvement include a normative statement about what is morally good? Not necessarily; it only includes the claim that an intervention is beneficial to the agent. When talking about moral enhancement it may be worth asking whether it is supposed to be beneficial for the person's moral aptitude or for their welfare more generally. The ambiguity of the phrase 'moral enhancement' can be partly explained by the presence of those two ways in which an intervention can be beneficial – beneficial to the 'moral profile' of the person or prudentially beneficial.

A similar question may arise for cognitive enhancement – we may be wondering whether an intervention is beneficial for a person's intellectual capacity or in the person's interest more generally speaking. There may be cases when cognitive enhancement is in a person's narrowly constructed interests, those related to their cognitive abilities and knowledge, but not necessarily in a widely constructed interest.

Let us consider two scenarios of cognitive enhancement. There may be cases when an improvement in certain cognitive abilities may not be in the overall interest of a person, for example, if there are substantial side effects that affect their physical or mental wellbeing. Take the example of Esperanza, a girl with borderline learning difficulties. Although after taking a new smart pill Esperanza becomes much better at mathematics and physics, yet the pill also happens to act directly on the neural circuitry involved in mood and emotion – causing Esperanza to feel depressed most of the time. Contrast this with the imagined story of Ernesto with a similar level of learning difficulty, who does not experience any obvious side-effects. However, after being able to learn so much more effectively, he becomes lonely – his old friends will not play with him anymore because now 'he is too smart', and neither will other pupils. Despite various efforts by teachers and parents over a long period of time, Ernesto becomes increasingly isolated. In Ernesto's case, although there are no straightforward adverse effects, an improvement in an aspect of cognitive function causes a behavioral change that brings net loss in wellbeing.

Consider another example, somewhat akin to that explored in Keyes' short story *Flowers for Algernon*. Esther has severe learning difficulties. She does not realize that she lacks certain capacities. She is a cheerful person and a pleasure to be around, and she enjoys her life. She does not display challenging behaviour and so does not require any additional medication. There is a new drug that was shown to improve cognitive function in people with less severe mental disability. Esther's mother decides to try this drug and there is a marked improvement of Esther's cognitive abilities. Unfortunately, the improvement is not

significant enough to allow her to be more independent, etc. – although she understands her environment better there is no great change in her well-being. However, for the first time, she starts to notice jokes that people make at her expense, and she has the acute awareness of her limitations. Although, as Mill famously wrote, it may be *"better to be Socrates dissatisfied than a fool satisfied"* (Mill, 1861/1991, p. 140), there will be some cases when the cost of knowledge or intelligent awareness is too high.

As Esther's example demonstrates, there may be some cases when intervention in cognitive capacities is beneficial in the narrow, but not in the wide sense. In this situation we could say that the intervention is an enhancement in the narrow sense, but is not an enhancement all-things-considered. The example of Esperanza and Ernesto demonstrates the importance of prediction and the estimation of cost and benefits. Despite the fact that cognitive enhancement *could* have such negative effects on Esperanza's and Ernesto's life, it can be argued that enhancement in such cases, (cognitive enhancement narrowly understood) usually brings more benefits than harms (in the wide sense) and is therefore worth pursuing. Education is important in our societies; high academic performance often translates into better career prospects and brings a number of other benefits. Thus, our experience with a number of instances of enhancement may lead us to say that cognitive enhancement (narrowly understood) is most likely to be in the person's interest.

A separate issue arises when considering enhancement as a beneficial intervention in the context of interventions in the moral sphere. If a similar logic is applied, intervention in the sphere of morality would only be an enhancement if it was beneficial for the subject of that intervention. Consequently, if we understand moral enhancement analogically to cognitive enhancement, there is nothing *prima facie* inconsistent in saying that moral enhancement can be going contrary to what is morally good. Moral enhancement thus understood may, for example, make people be less prone to act on moral reasons, give those reasons moral weight or act in a moral way. This is because moral enhancement will refer to intervention in the sphere of morality that brings an overall prudential benefit to an agent. Imagine Eric, who is deeply moved by moral considerations, strives for moral integrity and often acts on the basis of his moral beliefs. He spends a substantial amount of time on thinking about what is good and what is right, gives most of his disposable income to charities and spends many hours per week volunteering. However, his preoccupation with moral obligations has led to problems in family life. His wife threatens to leave him if he does not stop taking homeless people to their house and does not find at least some time to spend with her. In this case, acting as one think one ought to has negative consequences for Eric's overall wellbeing. Eric decides to strive to care less about others' misfortune.

Secondly, the 'moral' can refer to the overall ethical permissibility or obligatoriness of moral enhancement, and so be a more general reflection on the context of moral enhancement and the general consequences of it, and the judgment be all-things-considered. It can be argued that the second and first meaning of moral enhancement (making people better) and the all-things-considered moral assessment of this intervention could reasonably be collapsed into one. For example, one might argue that moral enhancement is an intervention that will result in people acting more morally, and the moral action will be the one that will maximize overall utility. At the same time, if the moral assessment of interventions is based also on maximizing utility, the two will likely coincide. The same is the case if one takes making people more

moral to mean being more virtuous and the ethical value of the enhancing intervention is judged on the basis of the extent to which it promotes virtue.

However, there are at least two situations when the two do not have to coincide. Firstly, the moral enhancement in the second sense could be achieved using morally reprehensible means, which in turn would influence our moral assessment of the intervention in question. If the moral enhancement of a given person could only be achieved at the price of a side-effect of inflicting strong pain on another person for a long time, we could reasonably argue that overall it was not worth it, even if enhancing the person would lead to them acting in a significantly more moral way after the intervention. Secondly, it is possible that making a person more concerned with morality may not have a positive overall effect under certain circumstances, for example that the subjective and the objective right[64] do not coincide and more consistent following what is subjectively right leads to morally worse outcomes. Another case can also happen – when an enhancing intervention would result in a less moral individual, but have overall good effects.

Some may object to the use of 'moral enhancement' as a phrase referring only to certain capacities and not carrying a clear normative message as well. Although it may be true that when we think about moral enhancement we automatically think about people being, morally speaking, better, this approach introduces confusion due to the 'moral' doing double, and sometimes triple, work: as a description of the target abilities that are improved, as a normatively loaded reference to whether that intervention results in people being morally better in some way, and as a reference to whether this enhancement is overall desirable from a moral point of view. This may introduce confusion when examining moral enhancement - we can be asking three questions to which we could give, at least in principle, diverging answers.

Moreover, if we find cognitive enhancement to be interesting, we are likely to find moral enhancement in the third sense (an enhancement of moral sphere beneficial to the agent) also interesting. It raises interesting questions about authenticity, free will and the moral nature of humans, the role of rational choice and of the motivation to be moral in choosing what moral stances one adopts. As a result, although most of the current discussion focuses on the second understanding of moral enhancement, the 'cognitive' understanding is interesting in its own right.

Hype or hope? - On the plausibility of moral enhancement

Against plausibility of moral enhancement

Persson and Savulescu (2008) in their paper *The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity* argue that non-traditional means of enhancement

[64] For subjective rightness, see Ross (1939), Carrit (1947) and Parfit (1988).

could contribute to the rising risk of considerable harm to a large number of people and, therefore, are ethically problematic (for criticisms of this claim see: Fenton, 2010; Harris, 2010). They suggest that this threat could be theoretically offset by moral enhancement. Accordingly, in their paper moral enhancement is seen as a potential tool for eliminating this alleged risk of large scale harm. Since small groups or even individuals could inflict serious harm, the aim of moral enhancement is to prevent the 'morally corrupt minority' from doing so (Persson & Savulescu, 2008).

Let us consider increasing empathy to illustrate some difficulties with this approach to moral enhancement. Increasing empathy comes to mind when thinking about what kind of intervention could carry out the task that Persson and Savulescu (2008) want moral enhancement to do. Lack of empathy is sometimes said to be correlated with criminal behaviour (Miller, 1988; Bush, Mullis, & Mullis, 2000; Kiehl, 2006) and it seems common-sense that an increased appreciation (be it cognitive, affective or both) of others' suffering would decrease the likelihood of behaviour that is likely to result in harm. For the purpose of this argument let us assume that there is an intervention that substantially increases empathy across different measures.

But increasing general empathy will most likely not be enough for several reasons, even when we assume that increased empathy is going to make a substantial difference in the motivation to act in a certain way. Firstly, we know that moral reasons are not the only basis for action and that prudential reasons can override moral ones. Thus, even if increased empathy would indeed give rise to reasons not to harm others that are stronger than before, those may not be enough to refrain from a fatally harmful action. As a result, it is reasonable to expect that even a highly efficient intervention will not be sufficient to abolish the possibility of harm completely.

Secondly, there may be cases when an increase in empathy would increase the risk of large scale harm. It is not clear that the allegedly morally corrupt minority that may pose a threat acts solely on the basis of non-moral reasons. This claim could be based on a conflation of two uses of 'moral' – one to describe a *kind* of reason for action and ethical assessment of actions. Thus, when we refer to 'morally corrupt minority' we mean 'those whose acts we judge as immoral'. But let us not forget the second meaning of 'moral' - it is possible for a terrorist to have her actions at least appearing to be based on moral reasons, that is, reasons of a moral *kind*. There is a long tradition of those claiming to be fighting for what they consider to be a better world and seeing inflicting harm as a necessary evil; sometimes we may support this struggle and sometimes we may denounce it (Merkl, 1986). We may disagree with the moral assessment that the terrorist has made, rejecting some or all of her reasons for action, disagree about which ends are desirable or simply disagree in our predictions of likely consequences – chances of success and the cost of bringing about the

desired end. On the other hand, there is also a long tradition of arguing against change despite the harms of the *status quo*.[65]

If we accept that empathy may provide a basis for some of those subjective moral reasons, it is also possible that there will be cases when empathizing with the suffering of the in-group will underlie the reasons for inflicting large-scale harm on the members of the out-group. Research suggests that although we are likely to empathize with the distress of any individual, empathy is vulnerable to the familiarity bias and here-and-now bias: people tend to favour family members, in-group members, close friends, people similar to themselves and those in physical proximity (Hoffman, 2000; Jones, 1991). If the increase in empathy for suffering of the in-group members is not offset by the increased empathy for the out-group, the increase in general empathy (other things being equal) could, in some circumstances, make the reasons for harming others stronger – for example, by adding poignancy to the observed suffering of those close to us.

An answer to this problem would be to selectively induce increased empathy towards those considered to be foes, and increase it to such an extent that other reasons are overridden in every situation that includes a large-scale harm. Combining such selectivity and force of an enhancing intervention seems indeed wildly implausible.

There are other reasons why moral enhancement may be unsuitable to serve the purpose Persson and Savulescu (2008) want it to serve. Persson and Savulescu (2008) seem to be focusing on wickedness as a cause of large-scale harm but, as Harris (2010) points out in his response in *Moral enhancement and freedom*, large-scale harm can be inflicted not only by 'the bad' but also by 'the mad'. Moreover, it can result from incompetence, stupidity, negligence and miscalculation (Rees in Harris, 2010). Thus, moral enhancement, even if possible and effective, is likely to be unable to offset the dangers allegedly brought by cognitive enhancement and the development of science in general.

Reconsidering our expectations

Moral enhancement for all?

I have outlined some of the reasons why moral enhancement may indeed not be able to eliminate the risk of large-scale harm, as Persson and Savulescu (2008) seem to require. But the expectation that moral enhancement *eliminate* the risk of large-scale harm seems to be not only potentially impossible (Harris 2010) but also unreasonably demanding.

One of the reasons why it is unreasonably demanding was pointed out by Harris (2010): the expectation that moral enhancement *ensure* safety by *eliminating* risk (Persson & Savulescu, 2008) seems

[65] As pointed out by Harris in his *Violence and Responsibility* (1980).

to be impossible to fulfill. Even if moral enhancement was fantastically efficient and cost-effective, it would be unlikely to eliminate the risk of large scale disaster because one malevolent person, who slipped through the net of enhancement or for whom the intervention did not work, is enough for the risk not to be eliminated. Even assuming that all disasters are caused only by malevolent individuals, the standard for the effectiveness of moral enhancing interventions is set very high. Although we may have such hope, we do not normally expect cognitively enhancing technologies to work for every single person, nor do we expect most very effective treatments to work in every single case.

Both cognitive and moral enhancement can be achieved by a variety of means. One of the possibilities is to use pharmacology. Pharmacological interventions often vary in effectiveness from individual to individual, and the influence of individual differences on outcomes is very well known, especially when the goal of intervention is to modify behaviour, mood or thinking processes. Predictions for outcomes (and side effects) for a single patient can be so unreliable that suitable medication is prescribed after a period of a trial-and-error search (Huskamp, 2003). Often, there is a group of patients that are unresponsive to any of the pharmacological remedies, and sometimes both to different types of medication and different types of therapies, as well as to combined approaches. The use of pharmacology that seems to be sensitive to individual differences and pharmacological interventions, at least as our experience so far suggests, is likely to work for some but not for others. The number of subjects who could experience a desired effect is likely to increase with the growing variety of available interventions, as new drugs and other technologies (such as deep brain stimulation (DBS)) are designed and tested to address the needs of those for whom nothing has yet worked (Mayberg et al., 2005; Berton and Nestler, 2006).

It is important to admit that pharmacological interventions have their limitations, but it is equally important not to forget about the cases when those interventions are effective. It may be regrettable that a drug is not effective for *all* (or even many), but denying the plausibility of moral enhancement because it does not work for all seems to be unjustifiably demanding.

Moreover, whether a potentially enhancing intervention is indeed enhancement depends on the context and on a particular person's needs. While weight gain may be an enhancement for an underweight individual, it would not be so for an obese person. Although we may think of a change as a typical cognitive enhancement, for example an increase in the ability to focus on a particular task while ignoring distractors (focusing attention), the same intervention may be neutral or even counterproductive in some tasks that require creativity (Pacholczyk & Harris, forthcoming).

Effectiveness

Another issue is the expected magnitude of change. The plausibility of moral enhancement could also be put in doubt if the interventions available seem to be insufficiently effective. If we adopted the goal of moral enhancement that Persson and Savulescu (2008) adopt, we would expect morally enhancing intervention to result in the overriding of any inclination or reason for inflicting large scale harm. Among 'the wicked' who would be willing to inflict large-scale harm, there are likely to be both those whose vector sum of reasons would be largely pointing towards causing a disaster, and others riddled with doubt but who in the

end decide to carry on. There will be opportunists who would change their minds with a little nudge, and those for whom bringing large-scale harm is a purpose of life and is consistent with at least some their strongly held beliefs. The impact of morally enhancing intervention would indeed have to be great to override all the subjective reasons for inflicting large-scale harm, strong attitudes and the impact of deeply engrained beliefs. With expectations about the effectiveness set so high, Persson and Savulescu (2008) are right to doubt that sufficiently effective moral enhancement will be possible. However, there is no reason why we should understand 'sufficient' as Persson and Savulescu (2008) do for the purpose of their paper.

I would like to briefly discuss another potential reservation. One could argue that since any moral enhancement possible in the near future will not be able to make *any* of the wicked people good, moral enhancement is implausible. This objection seems to be especially illustrative because of how it is misguided; it remains so *even if* we would agree with the factual premise. This is because, firstly, it misunderstands enhancement as necessarily bringing people from one extreme of the spectrum to another, and, secondly, it grossly oversimplifies the issue. We do not necessarily expect cognitive enhancers to turn stupid people into smart ones but rather to improve certain aspects of cognition, improve the ability to deal with certain kinds of tasks and so on. Although we may think about prototypically 'smart' and prototypically 'stupid' people, 'being smart' can mean many different things, and requires a whole range of cognitive processes. Secondly, 'smart drugs' do not make people smart or even smarter. They modify a narrow aspect of functioning that partly underlies abilities and behaviours that we then see as signs of being smart. Whether any particular case of modifying an aspect of cognitive functioning amounts to an enhancement can very well be context-specific. If cognitive enhancement does not make stupid people smart, why should we expect moral enhancement to turn wicked people into virtuous ones?

This is not to say that we could not *want* cognitive enhancers and moral enhancers to have this magnificent effect. We may hope for, imagine, and talk about the possibilities of radical human enhancement of a cognitive and moral realm, but even if radical enhancement is impossible, there is no reason why we should come to the conclusion that any enhancement is implausible. Naturally, we could point out that the effectiveness of enhancers is disappointing. It may indeed be the case that some of enhancing interventions will have such a small influence on functioning that, for most of us, they may not be worth the hassle. However, as some have pointed out, the cumulative and long-term impact of small changes can be substantial, yet are easily underestimated or altogether overlooked (Turner & Sahakian, 2006).

Another issue is that of comparative effectiveness as well as comparative cost-effectiveness. It may be the case that any pharmacologically induced change in moral functioning will be much less effective then more traditional means such as moral education. Even if that is the case, it may still be worth pursuing. It may be worth pursuing if pharmacological methods will be significantly cheaper or more accessible, and so ultimately cost-effective. Moreover, it may be the case that application of pharmacology and other novel means of enhancement may be more effective or cost-effective in certain specific circumstances, for specific groups or as a method that complements traditional means.

If we are going to consider pharmacological interventions that affect morality or cognition, there is no reason to be *automatically* discouraged if they have limited effectiveness (i.e., they do not turn maths idiots into maths geniuses or morally corrupt people into walking examples of virtue). We would likely find an

analogical threshold for efficiency impossible to reach for many pharmacological interventions. What we should rather look at is whether the effect of a single intervention or/and its repeated use is great enough, including the cumulative benefits of small beneficial changes to the extent possible, and look at comparative cost-effectiveness. It seems to be grossly premature to make strong judgments about those issues, given that neuroscientists only turned their interest to the cognitive science of morality a short time ago and that the empirical research on the effects of different interventions on the moral sphere of individuals is far from extensive. If we set our expectations to be comparable to those we have of cognitive enhancement and treatments for mental health disorders, the prospects of finding pharmacological or other moral enhancers seem to be much better.

I have argued that if we revisit our expectations about the effects of moral enhancement, it is premature to conclude that moral enhancement is implausible. I have addressed doubts about the plausibility of moral enhancements based on claims of its low effectiveness. Let us discuss the criticisms that suggest that moral enhancement is impossible even if it was possible to significantly modify relevant cognitive processes, motivations or emotions. In the next section I will briefly draw our attention to possible ways of approaching moral enhancement and discuss empirical research that can lead us to believe that using pharmacology to modify our moral sphere is not so far off.

What can we enhance? A multitude of potential targets

The ability to be moral is complex, so there are many substrates that can potentially be modulated. Our moral sphere involves the ability to make moral judgments, to be motivated by moral reasons, acting according to our moral beliefs, the ability to reflect on and critically analyze moral beliefs, and so on (consider the model presented by Rest, 1984; also reviewed in Bergman, 2004). Those tasks rest on a number of cognitive and affective capacities, which also can be modulated. However, knowing what is good is certainly not enough. The research on moral hypocrisy has demonstrated that although people often declare that a certain behavior is right, they often do not act on this belief – especially if it is against their interest to do so (Batson, Thompson, Seuferling, Whitney, & Strongman, 1999; Batson, 2008).

Moral behaviour encompasses a range of behaviours and includes refraining from doing what is wrong, doing what one ought to and doing good things beyond one's obligations. One of the ways in which behaviours could be modified is by influencing states that have a strong motivational component, such as emotions. Firstly, one could try to modulate other-oriented emotions which influence moral behaviour. Those include gratitude, awe, elevation (Silvers & Haidt, 2008), righteous anger (Rozin, Lowery, Imada, & Haidt, 1999), disgust and contempt. Until now, researchers have tended to concentrate on investigations of anger and disgust, and there is a significant (and growing) amount of research into the physiological underpinnings of those emotions and their neural correlates (e.g. Moll et al., 2005). Positively-valenced emotions such as gratitude have enjoyed less attention until recently (Watkins, Scheer, Ovnicek, & Kolts, 2006; Immordino-Yang, McColl, Damasio, & Damasio, 2009).

A number of self-conscious emotions (shame, guilt, embarrassment, and pride) play an important role in regulating behaviour, including moral behaviour. Emotions provide motivation to do good and to avoid

doing bad (Kroll & Egan, 2004). People can anticipate their likely emotional reactions to predicted behaviour based on previous experience and this prediction can influence moral choices (Tangney, Stuewig, & Mashek, 2007). Moreover, as is the case with a feeling of shame that is less central to morality, it can change the *post factum* behaviour. Research indicates that embarrassed people are inclined to behave in conciliatory ways in order to win approval from others (Cupach & Metts, 1990, 1992; Miller, 1996; Sharkey & Stafford, 1990). Some of those emotions may prove to be difficult to modify with novel means, but it is too early to make any strong conclusions on the issue. To show the possibilities that may emerge with more research, I will examine the effects on behaviour of one particular neurochemical – oxytocin.

A case study: Oxytocin

Oxytocin, a hormone and neurotransmitter originally known for its involvement in childbirth and lactation, has recently been shown to be involved in social behaviour. Studies in mice suggest that low levels of oxytocin correlate with impaired ability to recognise (Ferguson, Yung, Hearn, Matzuk, Insel, & Winslow, 2000; Ferguson, Aldag, Insel, & Young, 2001) and bond to one's peers (Winslow & Insel, 2002). These observations came in part from experiments with mice with a mutated oxytocin gene. Ferguson et al. (2001) showed that mouse knock-outs show a profound social recognition deficit despite normal olfactory and spatial learning abilities and that the social recognition ability can be fully restored by an injection of oxytocin in the medial amygdala. In contrast, a high level of this hormone seems to correlate with caring behaviour in rodents (Pedersen, Ascher, Monroe, & Prange, 1982).

In humans, oxytocin has also been shown to influence social behaviour and cognition. In humans oxytocin plays an important role in creating the mother-infant bond. Feldman and colleagues (2007) showed that a mother's level of oxytocin in the first trimester predicts the strength of the mother's attachment to the infant. Also, a boost to oxytocin levels in experimental settings commonly achieved by administration of a nasal spray, seems to promote trust (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005; Baumgartner, Heinrichs, Vonlanthen, Fischbacher, & Fehr, 2008) and generous behaviours (Barraza & Zak, 2009; Zak, Stanton, & Ahmadi, 2007). Oxytocin seems to influence social cognition (Theoridou, Rowe, Penton-Voak, & Rogers, 2009), increase some aspects of memory for social stimuli (Unkelbach, Guastella, & Forgas, 2008; Guastella, Mitchell, & Mathews, 2008) and to increase 'mind-reading' ability (Domes, Heinrichs, Michel, Berger, & Herpertz, 2007).

Substantial effect sizes obtained in the experiments on trust and generosity (for example, participants were 80% more generous in the oxytocin group than in the placebo group in the Zak et al. (2007) study and 30% more trusting in the Baumgartner et al. (2008) experiment), including some laboratory experiments that could have relatively high ecological validity (in Ditzen, Schaer, Gabriel, Bodenmann, Ehlert, & Heinrichs, 2008) couples were asked to argue and the frequency of positive behaviour such as listening, confirming or laughing during the conflict were measured) meant that the use of oxytocin in everyday life became more plausible.

Given decent effect sizes in experiments, we may worry that in many circumstances oxytocin could impede judgement and increase trust when trusting is unwarranted or even harmful (Damasio, 2005),

especially given the finding that oxytocin seems to restore trust after betrayals (Baumgarter et al., 2008). Based on this hypothesis, some have proposed the commercial and military application of oxytocin (Dethlefs, 2007). Those worries seem to be justified, given that although trust is an important social resource (Giddens, 1991), it can sometimes also be socially maladaptive (Greenspan, Loughlin, & Black, 2001). But is it indeed the case that an increase in oxytocin leads people to trust others indiscriminately?

Mikolajczyk and colleagues (2010a), suggest that the matter is somewhat more complex. They point out that in previous experiments, participants rarely met the same partner twice, nor had they clues that the person they interacted with was unreliable. Also, previous research suggested that the effects of oxytocin, for example on aggressive behaviour (this is especially well-illustrated in research on aggression in female rodents), are context dependent (Campbell, 2008; Pedersen, 2004). Mikolajczyk et al.'s (2010a) doubts were confirmed in research that used a customised economic trust game that allowed for repeated interaction with some partners being seemingly more trustworthy then others. Consistently with previous studies, they found that participants who received a nasal spray with oxytocin transferred more money to partners in comparison to participants in the control group. However, participants transferred more money to partners perceived as reliable but not to seemingly unreliable ones. This suggests that oxytocin administration does not increase trust when the partner appears unreliable. On the basis of these findings, Mikolajczyk et al. (2010a) proposed that the effect of oxytocin may be moderated by the perception of risk.

The effect of oxytocin on trust has to be confirmed for other contexts, although one recent study indicates that the effect is also present in circumstances that do not involve monetary transfers - participants who received oxytocin were 44 times more trusting that their privacy would not be violated than participants in the control condition (Mikolajczak, Pinon, Lane, de Timary, & Luminet, 2010). It is likely that oxytocin nevertheless will have some effect on the perception of how trustworthy others are. Yet let us make an assumption (reasonable on the basis of current evidence) that, generally speaking, the administration of oxytocin promotes trust – but it is unlikely that we would make some disastrous decisions because of this intervention. The research on the trust-promoting effects of oxytocin has an especially great potential application, given findings showing that we tend to underestimate people's trustworthiness (Fetchenhauer & Dunning, 2010) and because of the importance of trust for morally relevant actions.

Oxytocin also seems to improve empathy, but the effect was only prominent for less socially proficient participants – as measured by Autism Spectrum Quotient (AQ) – while there was no effect for the more socially proficient group (Bartz et al., 2010). Domes and others (2007) found that oxytocin improved performance only for difficult stimuli. These findings go against the tempting but simplistic view that oxytocin can be used as a universal prosocial enhancer that can turn all people into social-cognitive experts. Instead, perhaps unsurprisingly given our knowledge about the usual context-dependency and the impact of individual differences on the effects of both synthetic and naturally occurring pharmacological agents, oxytocin appears to help only some people. That is not to say that the effects are not substantial. In a Bartz et al. (2007) experiment, the administration of oxytocin equalised the differences in performance of low and high-performance groups in such a way that the performance of participants with high and low AQ scores was indistinguishable.

The exact mechanism underpinning the influence of oxytocin on trust, empathy and other potentially morally relevant abilities is still debated, but only looking at the known effects paints an intriguing picture. Assuming that the results discussed so far are confirmed, we may be able to gain a new, and possibly more convenient, means of influencing our level of trust and empathic ability, probably without a worry of overdoing it drastically (although there is some indication that oxytocin seems also to increase envy, see Shamay-Tsoory, Fischer, Dvash, Harari, Perach-Bllom, & Levkovitz, 2009). This may be not good enough for those envisaging radical human enhancement, but, as was argued in the 'reconsidering our expectations' section, the potential for this substantial, although not universal or radical enhancement, may be worth pursuing.

Although increases in trust or empathy may not eliminate (or maybe even decrease or keep constant) the risks of large scale harm, an increase in empathy for those from the lower-end of the functioning spectrum has the potential, via possible improvements in social cognition, to contribute to a better ability to take into consideration others' wellbeing – an ability fundamental to moral consideration. Positive behavioural effects demonstrated by couples during the conflict (which are most probably also mediated by an influence on amygdala and the level of stress) seem to point us towards a potential for a marked behavioural effect. Those can be especially useful in professions that require empathy, the well-developed ability to notice other's distress and maintaining prosocial attitudes also under stress and during a conflict, such as in the caring professions.

The ability to increase generosity and trust in the cases when the partner is judged to be trustworthy (or at least not particularly untrustworthy) seems highly unlikely to solve the serious political conflicts exacerbated by the lack of trust. However, we can easily imagine the situation when an increase in the frequency of acts typical of a Good or even a Splendid Samaritan – thus doing what is morally desirable but not obligatory – can have a notable and positive influence on what kind of social world we live in.

Doubts about moral enhancement – freedom and disagreement

In this section I will consider two objections to moral enhancement. I will first discuss the view that making people more moral is practically impossible due to the lack of consensus about what is and is not moral. Secondly, I will consider the view that moral enhancement would hurt the freedom of morally enhanced individuals (Harris, 2011).

Moral disagreement

Some may doubt the plausibility of moral enhancement, or even doubt the reasonableness of pursuing it, based on the claim that there is a substantial and possibly irreconcilable disagreement as to what is a moral way to go about things. If we disagree about what is moral, the argument might go, we cannot know which way we should modify our moral sphere – we do not even know what the goal of the modification should be!

There is at least one understanding of moral enhancement to which this doubt does not apply – moral enhancement understood as a modification in the sphere of moral functioning that is in the person's self-

interest outlined in the first section of this paper. Being more or less moral is not at issue here, and so the doubt does not have a bite.

And what about moral enhancement defined as making people more moral? It seems that the argument is in this case, at least *prima facie*, plausible. If moral disagreement undermines our moral knowledge, it could have consequences for the project of making people more moral, be it by non-traditional means and traditional ones such as moral education. But let us have a closer look.

Although there is a good amount of disagreement about what moral education should look like, most of us would not say for that reason that it is better to have no moral education at all, that we should not teach our children that lying is wrong or that striving to further develop ourselves as moral agents in our adulthood is a misguided proposition. Why is that? Firstly, because despite possible theoretical differences, there is a good amount of consensus about which acts or kinds of acts are morally wrong or morally right. On most, if not all, reasonable accounts of what is moral, killing a person for no other reason apart from the pleasure one derives from this act is wrong. There is also substantial agreement that, generally speaking, we ought to keep our promises or avoid lying. Moreover, there is a substantial amount of consensus about things that are necessary or conducive to moral agency and sensitivity, and conducive to morally good kinds of motivation, outcomes, etc. To give just one example – one of those things is concern and respect for other moral agents, which in turn requires a number of cognitive and affective capacities. The certain amount of agreement (more or less limited, depending on how high we will put the bar) means that the objection from disagreement does not apply to the numerous instances when disagreement is absent or weak enough. Objections from disagreement will not apply to improving our ability to be moral in those cases.

The presence of disagreement is often used to demonstrate the inadequacy of moral realism, and so to justify certain conclusions about the metaphysics of morals.[66] However, this argument is susceptible to the objection that it proves too much and - since it is an inference to the best explanation - the objection that there are alternative explanations of moral disagreement. When moral disagreement is present, it can be the result of several factors. It may be the result of disagreement about non-moral facts, both about morality and about the world. The disagreement can also have its source in some kinds of procedural failure in the reasoning process. Alternatively, the apparent disagreement may be an instance of the case when people are talking past each other, and do not understand each others' claims (Harman, 1975; Wong, 1984). In those cases we may hope that at least some disagreement may be removed. Abilities necessary to engage in a collective inquiry and discussion with others may be helpful in facilitating this process. Some kinds of enhancement (enhancement of reasoning skills, for example) may aid us in being better prepared for that process.

[66] For example, see Mackie's well-known 'argument from relativity'.

Some have assumed that all moral disagreement is in fact due to those reasons (e.g. Boyd, 1988), while others maintain that there are cases of moral disagreement between two people who are equally rational, and equally well informed about the non-moral facts and understand each others' claims (fundamental moral disagreement). Whether that is indeed the case seems to be a rather complex question and I will not attempt to give an answer here. However, *even if* we accepted the conclusion about the metaphysics of morals, it does not have straightforward implications for the possibility of moral actions and moral concern even in the case of fundamental moral disagreement. Why is that? It is because we cannot automatically get from the metaphysics of morals to the conclusion about moral knowledge and about what we should do. Let me just give one example of this – there are alternative metaphysical positions that have the potential to deal with the objections raised. It is possible, for example, to accept error theory and end up with moral fictionalism, where our make-believing in morality can be prudentially advisable (Joyce, 2005). Alternatively, moral non-cognitivists may seek to explain how the feelings, attitudes or prescriptions expressed in moral claims can be justified (see Hare (1981), Gibbard (1990) and Blackburn (1998) for theories of moral justification compatible with non-cognitivism). Those views can account for the apparent fundamental moral disagreement while leaving moral enhancement as a viable notion.

Although the most common, metaphysical arguments are not the only ones developed on the basis of the observation that moral disagreement exists. For example, McGrath (2007) defended an epistemological version of this argument. Epistemological arguments from disagreement seek to undermine moral knowledge by showing that regardless of the metaphysics of moral facts, we can reasonably expect to have much less moral knowledge that we previously thought. Consider the following passage from Sidgwick's *The Methods of Ethics*:

> [I]f I find any of my judgments, intuitive or inferential, in direct conflict with a judgment of some other mind, there must be error somewhere: and if I have no more reason to suspect error in the other mind than in my own, reflective comparison between the two judgments necessarily reduces me temporarily to a state of neutrality. (Sidgwick, 1907/1981, p. 342).

McGrath (2007) develops a parallel argument that applies not to certainty, but rather to moral knowledge. When moral beliefs are subject to disagreement and Sidgwick's condition is satisfied (that is, if one has no more reason to suspect that the other person is mistaken than that it is oneself who erred), one is not holding knowledge about the contested issue; and that is the case even if the belief happens to be true. In fact, McGrath (2007) develops a stronger version of this claim by arguing that all controversial moral issues (such contentious matters in applied ethics and culture) fulfil Sidgwick's condition; let us accept this last claim for the purposes of the argument. What consequences does it have for the moral enhancement project?

The consequences are far from straightforward. In those cases it does not follow that we should abandon, prohibit or find moral enhancement an untenable proposition – and that applies to both moral education and other non-traditional means of enhancement. Firstly, in cases that apparently satisfy Sidgwick's condition we may still have some problems with justifying why exactly it is rational for us to trust others' moral intuitions as much as we trust ours, and why, as a consequence, we should abandon our belief

(Wedgwood, 2010). But let us assume that some version of Sidgwick's proposal applies and so in many cases of controversy it is rational for us to abandon our beliefs.

Non-traditional moral enhancement is unlikely to be specific enough to change the moral appraisal of any particular controversial issue. It is more likely to slightly modify some propensities to react, perceive and behave by increasing impulse control, empathy, trust or reducing fear responses and so on. Naturally, we can still disagree about issues such as whether a higher level of trust is conducive to moral outcomes. However, if we accept Sidgwick's (1907/1981) advice to hold our judgements we are still left with the question 'so what should we do now?' Let us say that we disagree about whether Jane should increase, decrease or maintain her empathic ability (we fundamentally disagree about all three possibilities). What behaviour would constitute holding our judgement on this issue? Some may say that we should leave things as they are. But there is no reason why we should privilege the *status quo* option over other possibilities, given that there is disagreement also about the *status quo*. Thus, moral disagreement is problematic as a support for leaving things as they are. We are still faced with the question 'what should we do next?' The answer could be that it is only rational for us to have *no moral views* at all on the contentious matter and use other reasons to decide on the course of action.

It is important to remember that we have developed political means of dealing with moral disagreement and sometimes find disagreement to be a constructive force necessary for change. In liberal societies moral education is often about developing the ability of persons to be autonomous moral agents, providing them the possibility of gaining reasoning skills and exposure to moral problems to aid this development. We tend to protect the freedom of people to disagree with commonly held views. We also have political frameworks that aid us in dealing with moral disagreement and often seek the state to be as neutral about issues of morality as it is possible. We tend to protect the private sphere – the freedom of parents to raise their children as they see fit is interfered with only in cases of clear parental failure; we struggle to protect freedom of conscience, and so on. We accept that people have different ideas about what a good life is about and value the ability of individuals to act consistently with their idea of the good life and morality, and, generally speaking, restrain this possibility only when we have strong justification for doing so. Even given the doubts that an agent may have about what is right, we are likely to find the adoption of a moral stance (for example, as opposed to narrowly self-interested stance) to be valuable.

One could argue that the possibility of moral enhancement in this liberal framework would be likely to deepen the disagreement - which could be seen as undesirable prudentially or morally speaking. We may therefore have good reasons to make people less bothered about morality in cases when disagreement arises (this would be a solution consistent with the view that it is rational for us to abandon our belief in certain cases of disagreement). Interestingly, an argument for making people suspend their judgment and not act motivated by moral reasons under these particular circumstances is an argument for a certain kind of enhancement. If one supported this argument using *moral* reasons this would be an argument for a specific kind of moral enhancement understood as making people more moral. If the rationale is prudential, we have a case for prudentially beneficial intervention into our moral sphere.

To sum up, moral disagreement has much less straightforward consequences for moral enhancement that we may have thought at the outset. Firstly, it only applies to moral enhancement understood as making

people more moral, and not to moral enhancement as a prudentially beneficial modification of the moral sphere. Also, it does not apply to a whole array of issues that we tend to agree on, including the issues of what is conducive to morally desirable moral sentiments, motivations, outcomes, etc. If we treat moral disagreement as giving rise to a valid and strong argument against certain views on the metaphysics of morals, there is still much explaining to be done of what impact it should have on our moral knowledge and subsequent actions. We can, for example, adopt a non-realist view of morality that is not susceptible to the objection from disagreement and work from there. What we have learned from the discussion on the possible sources of disagreement is that disagreement about non-moral facts, procedural failure, bias and lack of proper discussion can all give raise to disagreement about moral issues. We may therefore have a good reason to support both traditional and non-traditional means of improvement that would aid us in dealing with those disagreements better then we now do.

It is also unclear how an epistemological argument from disagreement should impact our behaviour, but it is unlikely to support the *status quo*. If we indeed think that holding our judgement means abandoning moral considerations in controversial cases and that this is what we *ought to* do, we may have a good case for a particular kind of moral enhancement. Also, let us not forget that we have political means of dealing with moral disagreement. Respecting moral agents' decisions and allowing moral agents to pursue their idea of the good in the private sphere (and discuss and argue for it in the public sphere) is one of them. Unless we have other strong reasons to treat non-traditional moral enhancement differently, this also applies to those cases of moral enhancement.

Freedom

State coercion and freedom

There are two main ways in which we may think about moral enhancement as threatening freedom. Firstly, moral enhancement may be imposed by governments. Moral enhancement would make a person better in some way but it is carried out against this person's will or without their knowledge. Secondly, even if the person consents to or chooses to undergo a morally enhancing intervention, an intervention in the moral sphere may be seen as diminishing that person's ability to make moral choices. Let me start out by briefly addressing the first objection.

The concern that states could use novel technological means to manufacture consent, dissolve dissent and surreptitiously force citizens readily springs to mind – those who deal with the ethics of enhancement are always reminded of the 'Brave New World' scenarios. In the discussions following the boom in antidepressant prescriptions many commentators compared Prozac and other such drugs to *soma* (for utopian and dystopian takes on pharmacology see Schermer, 2007). Although after years of antidepressants' presence on the market the threats envisaged by the most pessimistic and imaginative commentators have hardly turned into reality, those very often rhetorical and alarmist evocations of dystopias remind us about a number of important political issues surrounding the conditions for a workable democratic

state (including the importance of dissent), the potential threats of a strong security-state to individual freedom and the limits of justified paternalism.

However, those commentators *raise questions* rather than provide answers and we should be wary of a knee-jerk response against the mandatory use of moral enhancement. It is generally accepted that we relinquish some aspects of our freedom in exchange for security or other benefits of living in a society. It *could* be – in some circumstances – justified for the government to impose or strongly encourage morally enhancing interventions. For example, it seems to be prerogative of the parents, but also lay within the interest of the state, to make sure that children acquire and regulate their behaviour according to the rules of co-existence. The failure of parents to install those rules may provoke the intervention of the state (by taking children into foster care or mandating contact with a social worker), especially if this failure translates into highly disruptive or illegal behaviour.

In many modern democracies it is currently legal to use pharmacological interventions for some offenders and some people with mental health problems; the use of those interventions can be both compulsory and voluntary. Those measures include mental health treatment for individuals that are judged to be posing a danger to themselves or others and chemical castration of sex offenders (Harrison, 2007; Grubin & Beech, 2010).The legislation that regulates the uses of those measures surely evokes legal and ethical controversies. However, those examples illustrate the point that what we need is a discussion about what constitutes a legitimate justification for employment of those techniques. This includes questions about justification of paternalism and interference with individual freedom in order to prevent harm to others, the boundary between the public and the private sphere and so on. It is not within the scope of this paper to address those questions adequately, but it is worth noting that the appraisal of moral enhancement imposed by governments will often significantly depend on our ideas about paternalism and the conditions for the legitimacy of state interference. Also, those questions were addressed at length and hotly debated specifically in the context of the use of pharmacology by scholars dealing with ethical issues in mental health, medical and criminal law, and we should remember to take advantage of those debates.

Voluntary moral enhancement and freedom

Are there cases when voluntary modification of the moral sphere would decrease our freedom in an undesirable way? I would like to focus on one way in which such loss of freedom could be thought of – the worry that moral enhancement, if successful, would eradicate the possibility of moral life by eliminating choice.

Harris (2011) starts his discussion of moral enhancement with a passage from Milton's *Paradise Lost*. In this passage Milton's God states that if man surrenders to Satan's temptation, he has only himself to blame as 'I made him just and right, sufficient to have stood and free to fall.' As Harris (2011) correctly points out, millennia of evolution resulted in us – creatures with vigorous moral sentiments, a sense of justice and right. Some may be tempted to say that since we are endowed with the capacity for moral responsibility and moral life, we do not need to pursue specifically *moral* enhancement. This, however, would be a rather obviously premature conclusion. The given ability or predisposition for moral life is normally strengthened

and modified by the influence of culture – bringing up children to know right from wrong, developing the ability for moral reflection and discussion and transmitting knowledge about non-moral facts relevant for our moral choices, etc. We not only accept, but usually also support and encourage traditional means of moral development and improvement, including moral education and reflection.

Moreover, there is no reason to suppose that evolution has equipped us well enough to deal with life in the modern society – in increasingly complex social and political circumstances and with the plethora of new ways of inflicting harm that we came up with over the last couple of centuries. Indeed, the results of empirical research give us good reasons to suspect that in some respects we are ill equipped to deal with problems that are facing us today, and it is especially vivid in the case of moral appraisal of indirect action (see Paharia, Kassam, Greene, & Bazerman, 2009). For example, one study on moral intuitions suggests that harm involving physical contact is often judged to be morally worse than harm without such contact, despite the fact that when subjects are asked about moral relevance of physical contact, they often deny that it is relevant (Cushman, Young, & Hauser, 2006).

What worries Harris (2011) seems to be that the current discussion has focused on moral enhancement as eliminating characteristics or dispositions that are *often* conducive to wrongdoing. The problem is that the same sorts of characteristics seem to be necessary *"not only for virtue but for any kind of moral life at all"* (Harris, 2011, p. 3). Harris (2011) makes several largely independent arguments, but I will only focus on a claim most relevant to our discussion about freedom. And this is the statement that *there is no virtue in doing what you must*.

Even without getting deep into the centuries-long discussion about determinism and free will, and what we exactly mean by choice, Milton's words cited at the beginning of this section bring to our attention an important pre-condition for moral responsibility. Moral responsibility, and the virtuous choosing to do what is good and right, necessitates both the possibility to fall and the freedom to choose to fall (Harris, 2011). It seems that there are two necessary ingredients for a 'freedom to fall', understood as such, to be realized. One, we have to have the ability to choose (whatever we mean by it), and, secondly, we have to be presented with a situation where there indeed is a choice. How can moral enhancement be a threat? One way is when the temptation (or, in other words, the reasons or motivation to do wrong) is eradicated. Moral enhancement as proposed by Douglas (2008) and Persson and Savulescu (2008) may be worrying because it would modify our moral sphere in such a way that there would be no scope for choice. There is no virtue, the argument may go, in overcoming our racial prejudice and not acting on it, if there is no prejudice, no virtue in staying calm and composed, when we cease to feel anger or fear.

There certainly is a virtue in overcoming unjustified prejudice and not acting on it. Most of us, however, would happily forgo many possibilities of being virtuous if only we could minimize the harm caused by this prejudice when we fail to be virtuous enough to stop ourselves from acting on it. Also, there are plenty of temptations we *do not have*. We usually are not tempted to steal toys from children in the playground, to shave off our partner's hair while they are asleep or poison our neighbours. Even if we could, we are unlikely to want to create those temptations only so that we can overcome them – and we do not find our moral lives impoverished because of lack of those temptations. Those who value the possibility of exercising virtue

should not worry. Moral enhancement, even if more efficient than we can reasonably predict it to be in the coming decades, is unlikely to dispose of situations that require virtuous behaviour.

The statement that *there is no virtue in doing what you must* could, in fact, be evoked in support of moral enhancement understood as modification of our moral sphere, whether it would result in a more or less moral person. In some cases, we have a limited choice about whether or not to be moved by moral considerations (Narvaez & Lapsley, 2009). We naturally can, both for the better and for worse, use reflection to moderate their influence, work at changing our moral intuitions, adopt philosophies that will help to offset unwanted influence, etc. But conscious control over our moral lives seems to be limited. If to be moral we need to be able to choose to fall, this possibility can be greatly improved by modification of our moral sentiments and related first-order desires. But maybe it is this freedom that we fear?

One could make a consequentialist argument and say that it is very likely that having the freedom not to be morally facilitated, for example, by the increased ability to silence emotions such as guilt after wrongdoing, will eventually lead to the world being generally worse. This is, however, a different kind of argument from the worry about impairment to freedom. It is an argument for restricting freedom, and not for letting people decide how to lead their lives.

Our capacity to be moral and the particular values that inform the expression of this ability seems, as many other things of our life, to be determined by innate inclinations and early experience, including moral education. This does not preclude further moral development or change. Many of us actively modify our moral sphere and rethink the rules and values we internalized during the course of our lives. Some of this modification occurs without extensive reflection and is a response to the changes in our knowledge or social circumstances. In other cases we think about our values, our morally relevant automatic reactions, find inconsistencies or faults in our moral reactions and then either try not to act on them or to change them. If in this process we use pharmacological 'helpers', it does not seem to essentially change the process; it simply increases the means available to us.[67] Pharmacological means are unlikely to replace traditional means of moral development, but may be well placed to complement them.

Conclusion

In this paper, I have explored the idea of moral enhancement and looked at three possible interpretations of this phrase that are worth keeping in mind when discussing moral enhancement. I have challenged the assumption that 'moral enhancement' necessarily means making people morally better and suggested that a non-normative understanding is also interesting. Contrary to some commentators, I have

[67] I will assume that there is nothing in principle wrong with using chemical means for enhancement. For an enlightening discussion of that subject please see Harris (2007).

suggested that the moderately effective pharmacological interventions of this kind are already with us, and discussed oxytocin as an example of a potential moral enhancer.

The discussion of the objection to the use of moral enhancement revealed that the consequences of moral disagreement and the worries about decreasing freedom are more problematic that it might have seemed. Metaphysical arguments from disagreement are not strong enough to convince us that moral enhancement understood as making people better is untenable, and epistemological arguments do not support the *status quo*. In fact, if we find epistemological arguments from disagreement convincing, it may lead us to support certain modifications in our moral sphere. Similarly, the worries about the effect of moral enhancement on freedom seem to, at the very least, merit more discussion. Discussed doubts about the voluntary use of moral enhancement seem not to be overly persuasive and the concern about the use of morally modifying interventions by the state does not apply to moral enhancement only, but is better understood as a wider political question about the criteria for the legitimacy of state interference.

References

Barraza, J. A., & Zak, P. J. (2009). Empathy toward strangers triggers oxytocin release and subsequent generosity. *Annals of the New York Academy of Sciences, 1167(1),* 182-189.

Bartz, J. A., Zaki, J., Bolger, N., Hollander, E., Ludwig, N. N., Kolevzon, A., & Ochsner, K. N. (2010). Oxytocin selectively improves empathic accuracy. *Psychological Science, 21(10),* 1426-1428.

Batson, C. (2008). Moral masquerades: experimental exploration of the nature of moral motivation. *Phenomenology and the Cognitive Sciences, 7(1),* 51-66.

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: appearing moral to oneself without being so. *Journal of Personality and Social Psychology, 77(3),* 525-537.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., & Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron, 58*, 639-650.

Bergman, R. (2004). Identity as motivation: Toward a theory of the moral self. In D. K. Lapsley & D. Narvaez (Eds.), *Moral Development, Self, and Identity* (21-46). Mahwah, NJ: Erlbaum.

Berton, O., & Nestler, E. J. (2006). New approaches to antidepressant drug discovery: beyond monoamines. *Nature Reviews Neuroscience, 7(2),* 137-151.

Blackburn, S. (1998). *Ruling Passions*. Oxford: Clarendon Press.

Boyd, R. N. (1988). How to be a moral realist. In G. Sayre-McCord (Ed.), *Essays on Moral Realism* (181-228). Ithaca and London: Cornell University Press.

Bush, C. A., Mullis, R. L., & Mullis, A. K. (2000). Differences between offender and nonoffender youth. *Journal of Youth and Adolescence, 29,* 467-478.

Carrit, E. F. (1947). *Ethical and political thinking*. Oxford: Clarendon Press.

Campbell, A. (2008). Attachment, aggression and affiliation: The role of oxytocin in female social behavior. *Biological Psychology, 77,* 1-10.

Cupach, W. R., & Metts, S. (1990). Remedial processes in embarrassing predicaments. In J. Anderson (Ed.), *Communication Yearbook* (323-352). Newbury Park, CA: Sage.

Cupach, W.R., & Metts, S. (1992). The effects of type of predicament and embarrassability on remedial responses to embarrassing situations. *Communication Quarterly, 40(2)*, 149-161.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgments: testing three principles of harm. *Psychological Science, 17,* 1082-1089.

Damasio, A. (2005). Brain trust. *Nature, 435*, 571.

Dethlefs, D. R. (2007). Chemically enhanced trust: Potential law enforcement and military applications for oxytocin (Naval Postgraduate School Master thesis). Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA475794&Location=U2&doc=GetTRDoc.pdf

Ditzen, B., Schaer, M., Gabriel, B., Bodenmann, G., Ehlert, U., & Heinrichs, M. (2009). Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict. *Biological Psychiatry, 65,* 728-731.

Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. C. (2007). Oxytocin improves “mind-reading” in humans. *Biological Psychiatry, 61,* 731-733.

Douglas, T. (2008). Moral enhancement. *Journal of Applied Philosophy, 25(3),* 228-245.

Eyler, J. M. (2003). Smallpox in history: the birth, death, and impact of a dread disease. *The Journal of Laboratory and Clinical Medicine, 142(4),* 216-220.

Feldman, R., Weller, A., Zagoory-Sharon, O., & Levine, A. (2007). Evidence for a neuroendocrinological foundation of human affiliation: plasma oxytocin levels across pregnancy and the postpartum period predict mother–infant bonding. *Psychological Science, 18*, 965-970.

Ferguson, J. N., Yung, L. J., Hearn, E. F., Matzuk, M. M., Insel, T. R., & Winslow, J. T. (2000). Social amnesia in mice lacking the oxytocin gene. *Nature Genetics, 25*, 284-288.

Ferguson, J.N., Aldag, J.M., Insel, T.R., & Young, L.J. (2001). Oxytocin in the medial amygdale is essential for social recognition in the mouse. *Journal of Neuroscience, 21*, 8278-8285.

Fetchenhauer, D., & Dunning, D. (2010). Why so cynical? *Psychological Science, 21(2),* 189-193.

Gibbard, A. (1990). *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

Giddens, A. (1991). *Modernity and self-identity*. Cambridge, MA: Polity.

Greenspan, S., Loughlin, G., & Black, R. S. (2001). Credulity and gullibility in persons with mental retardation: A framework for future research. *International review of research in mental retardation, 24*, 101-135.

Grubin, D., & Beech, A. (2010). Chemical castration for sex offenders. *British Medical Journal, 340*, c74.

Guastella, A. J., Mitchell, P. B., & Mathews, F. (2008). Oxytocin enhances the encoding of positive social memories in humans. *Biological Psychiatry, 64,* 256-258.

Hare, R. M., (1952). *The Language of Morals*. Oxford: Clarendon Press.

Harman, G. (1975). Moral relativism defended. *Philosophical Review, 84*, 3-22.

Harris, J. (1980). *Violence and Responsibility*. London: Routledge & Kegan Paul.

Harris, J. (2007). *Enhancing Evolution.* Princeton, N.J., Woodstock: Princeton University Press.

Harris, J. (2007). *Enhancing evolution: the ethical case for making better people*. Princeton, N.J., Woodstock: Princeton University Press.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics, 25(2),* 102-111.

Harrison, K. (2007). The high-risk sex offender strategy in England and Wales: is chemical castration an option? *The Howard Journal of Criminal Justice, 46(1),* 16-31.

Hoffman, M. L. (2000). *Empathy and moral development: implications for caring and justice*. Cambridge, UK: Cambridge University Press.

Huskamp, H. A. (2003). Managing psychotropic drug costs: will formularies work? *Health Affairs, 22(5),* 84-96.

Immordino-Yang, M. H., McColl, A., Damasio, H., & Damasio, A. (2009). Neural correlates of admiration and compassion. *Proceedings of the National Academy of Sciences, 106(19),* 8021-8026.

Jones, T. M. (1991). Ethical decision making by individuals in organizations: an issue-contingent model. *Academy of Management Review, 16,* 366-395.

Joyce, R. (2005). Moral Fictionalism. In M. E. Kalderon (Ed.), *Fictionalism in metaphysics* (287-313). Oxford: Oxford University Press.

Kiehl, K. A. (2006). A cognitive neuroscience perspective on psychopathy: evidence for paralimbic system dysfunction. *Psychiatry Research, 142,* 107-128.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature, 435*, 673-676.

Mayberg, H. S., Lozano, A. M., Voon, V., McNeely, H. E., Seminowicz, D., Hamani, C., Schwalb, J. M., & Kennedy, S. H. (2005). Deep brain stimulation for treatment-resistant depression. *Neuron, 45(5),* 651-660.

McGrath, S. (2007). Moral disagreement and moral expertise. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics Vol. 4* (87-107). Oxford: Oxford University Press.

Merkl, P. H. (1986). *Political violence and terror: motifs and motivations*. Berkeley; London: University of California Press.

Mikolajczak, M., Gross, J. J., Lane, A., Corneille, O., de Timary, P., & Luminet, O. (2010a). Oxytocin makes us trusting but not gullible. *Psychological Science*, *22*, 1072-1074.

Mikolajczak, M., Pinon, N., Lane, A., de Timary, P., & Luminet, O. (2010). Oxytocin not only increases trust when money is at stake, but also when confidential information is in the balance. *Biological Psychology, 85(1),* 182-184.

Mill, J. S. (1859/1991). *On liberty and other essays*. New York: Oxford University Press.

Miller, R. S. (1996). *Embarrassment. Poise and peril in everyday life*. New York: Guilford Press.

Miller, P. A., & Eisenberg, N. (1988). The relation of empathy to aggressive and externalizing/antisocial behavior. *Psychological Bulletin, 103(3),* 324-344.

Moll, J., de Oliveira-Souza, R., Moll, F. T., Ignacio, F. A., Bramati, I. E., Caparelli-Daquer, E. M., & Eslinger, P. J. (2005). The moral affiliations of disgust: A functional MRI study. *Cognitive and Behavioral Neurology, 18(1),* 68-78.

Pacholczyk, A., & Harris, J. (forthcoming). Dignity and enhancement. In N. J. Palpant & S. C. Dilley (Eds.), *Human Dignity in Bioethics*.

Paharia, N., Kassam, K. S., Greene, J. D., & Bazerman, M. H. (2009). Dirty work, clean hands: The moral psychology of indirect agency. *Organizational Behavior and Human Decision Processes, 109(2),* 134-141.

Parfit, D. (1988). What we together do. Retrieved from: http://individual.utoronto.ca/stafforini/parfit/parfit_-_what_we_together_do.pdf

Parfit, D. (1997), Reasons and motivation. *Proceedings of the Aristotelian Society, Supplementary Volume, 71,* 99-130.

Pedersen, C. A., Ascher, J. A., Monroe, Y. L., & Prange, A. J. (1982). Oxytocin induces maternal behavior in virgin female rats. *Science, 216*, 648-650.

Pedersen, C. A. (2004). Biological aspects of social bonding and the roots of human violence. *Annals Of The New York Academy Of Sciences, 1036*, 106-127.

Persson, I., & Savulescu, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of Applied Philosophy, 25(3),* 162-177.

Rest, J. R. (1984). The major components of morality. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Morality, Moral Behavior, and Moral Development* (24-37). New York: Wiley.

Ross, D. W. (1939). *The foundations of ethics*. Oxford: Clarendon Press.

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology, 50*, 703-712.

Schermer, M. (2007). "Brave New World' versus 'Island' - Utopian and dystopian views on psychopharmacology. *Medicine, Health Care and Philosophy, 10(2),* 119-128.

Shamay-Tsoory, S. G., Fischer, M., Dvash, J., Harari, H., Perach-Bllom, M., & Levkovitz, Y. (2009). Intranasal administration of oxytocin increases envy and schadenfreude (gloating). *Biological Psychiatry, 66(9),* 864-870.

Sharkey, W. F., & Stafford, L. (1990). Responses to embarrassment. *Human Communication Research, 17(2)*, 315-335.

Sidgwick, H. (1907/1981). *The methods of ethics.* Indianapolis: Hackett.

Silvers, J. A., & Haidt, J. (2008). Moral elevation can induce nursing. *Emotion, 8(2),* 291-295.

Tangney, J. P., Stuewig, J., & Mashek, D.J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58*, 345-372.

Theoridou, A., Rowe, A. C., Penton-Voak, I. S., & Rogers, P. J. (2009). Oxytocin and social perception: Oxytocin increases perceived facial trustworthiness and attractiveness. *Hormones and Behavior, 56,* 128-132.

Turner, D., & Sahakian, B. (2006). Ethical questions in functional neuroimaging and cognitive enhancement. *Praxis: International Journal of Technology Assessment and Ethics of Science, 4(2),* 81-94.

Unkelbach, C., Guastella, A. J., & Forgas, J. P. (2008). Oxytocin selectively facilitates recognition of positive sex and relationship words. *Psychological Science, 19*, 1092-1094.

Watkins, P., Scheer, J., Ovnicek, J., & Kolts, R. (2006). The debt of gratitude: Dissociating gratitude and indebtedness. *Cognition & Emotion, 20(2),* 217-241.

Wedgwood, R. (2010). The Moral Evil Demons. In R. Feldman & T. Warfield (Eds.), *Disagreement* (216-246). Oxford: Clarendon Press.

Winslow, J. T., & Insel, T. R. (2002). The social deficits of the oxytocin knockout mouse. *Neuropeptides, 36,* 221-229.

Wong, D. (1984). *Moral relativity.* Berkeley, CA: University of California Press.

Zak, P. J., Stanton, A. A., & Ahmadi, S. (2007). Oxytocin increases generosity in humans. *PLoS One, 2 (11),* e1128.

Chapter 8
Free will, punishment and neurotechnologies

Elizabeth Shaw
University of Edinburgh
School of Law
✉ E.Shaw-2@sms.ed.ac.uk

Abstract This paper discusses the issue of employing neurotechnologies in the penal system to 'morally enhance' offenders. Some rehabilitation programmes (e.g. for drug addicts and sex offenders) currently aim to reduce the strength of offenders' deviant urges and to increase their control over these urges. However, more radical interventions may become technically possible, e.g. neurological interventions that aim to alter the offender's goals and values to something more morally acceptable. Both consequentialists and retributivists may oppose such radical methods. Retributivists might argue that only traditional punishment respects the offender's 'free will'. This response is open to a number of challenges. For instance, many punishment theorists (e.g. the retributivist, Michael Moore) consider it plausible that human conduct is determined by factors outwith the agent's control. But if this is correct, why should behaviour count as 'free' when it is determined by certain factors (e.g. genes and upbringing) and not by other factors (e.g. neurological intervention)? Consequentialists may emphasise the potential for neurotechnologies to be abused. For instance, the state might attempt to manipulate law-breakers who have legitimate political grievances. However, it is unclear whether, from a purely consequentialist perspective, the risk of abuse will always outweigh the good consequences to be achieved by reshaping the uncontroversially abhorrent values of certain offenders, e.g. racist extremists. My paper argues against attempting to alter offenders' goals and values using neurotechnologies that wholly or largely circumvent the offender's rationality. My approach aims to be compatible with determinism and to avoid the problems of consequentialism. It appeals to the values of equality and moral dialogue. It opposes types of intervention, which would create an immense inequality of power between state officials and other citizens, arguing that the state does not have the *moral status* to use such methods. Law-breakers are still part of the moral community and efforts to alter their values should be via *relationships* with others (e.g. through victim-offender mediation), rather than by direct neurological interventions.

Keywords determinism, punishment, moral enhancement, equality

Introduction

This paper discusses the issue of employing neurotechnologies in the penal system to 'morally enhance' offenders. Some rehabilitation programmes (e.g. for drug addicts and sex offenders) currently aim to reduce the strength of offenders' deviant urges and to increase their control over these urges. However, more radical interventions may become technically possible, e.g. neurological interventions that aim to

directly alter the offender's goals, or values to something more morally acceptable. Henceforth, neurological interventions which aim to alter such features of agency will be referred to as 'direct interventions'.

Such radical interventions are intuitively troubling. One kind of objection to these interventions cites the bad consequences that might flow from their use or abuse. Even well-intentioned use of neurotechnologies within the penal system might have unintended harmful results. An intervention might fail to achieve its purpose. For instance, a technique designed to 'improve the moral character' of the offender might have the opposite effect. Or, even if it does achieve its purpose it might do so at a significant cost. For example, the intervention may have side-effects which are harmful to the offender's health. Furthermore, there is a risk that the authorities will abuse their power. For instance, the state might attempt to manipulate lawbreakers who have legitimate political grievances. However, from a purely consequentialist perspective, it seems that all of these risks have to be factored into the utilitarian calculation and weighed against the potential benefits of direct interventions. It is far from obvious that these risks will in all cases necessarily outweigh the potential benefits (e.g. in terms of crime reduction and protecting society).

A seemingly obvious objection to employing direct neurological interventions to alter offenders' motivations is that these interventions threaten to undermine the offender's 'free will'. The first section of this paper will argue that free-will based objections to direct interventions are actually far from straightforward. For instance, many punishment theorists consider it plausible that human conduct is determined by factors outwith the agent's control (e.g., Moore, 1985). But if this is correct, it becomes difficult to explain why behaviour should count as 'free' when it is determined by certain factors (e.g. genes and upbringing) and not when it is determined by other factors (e.g. neurological intervention).

My paper adopts an approach which is not based on the notion that direct interventions necessarily violate free will. It also aims to avoid the problems of consequentialism. It presents an objection to certain types of direct intervention, which appeals to the values of equality and moral dialogue. This paper does not oppose the use of *all* neurological interventions within the criminal justice system. It specifically focuses on the idea of using neurotechnologies to deal with offenders who are basically rational, in an attempt to directly re-shape these offenders' values or goals. It argues that such interventions are unacceptable because they would exclude offenders from the moral community, and because the state does not have the *moral status* to use such methods.

Key terms

It is important at the outset to clarify some key terms. Firstly, the thesis of 'Determinism' implies that every person's actions are caused by earlier events whose occurrence was not under the agent's control. The sense of 'cause' that is being used here does not merely refer to the presence of *necessary conditions* for the occurrence of the action. Nor does it mean that prior events just made the action *more likely* to occur. Rather, it means that prior events (taken together with the laws of nature) were *causally sufficient* to produce that action – i.e., given the occurrence of the earlier events, the action *had to occur*, no matter what else may have been the case. Determinism does not mean that our actions do not affect what happens to us (fatalism). Nor does it mean that mental events (e.g. our intentions, decisions and desires) do not cause our

actions. Rather, the thesis of determinism implies that, if our actions are caused by mental events, then those mental events were themselves produced by prior events that were causally sufficient for the occurrence of those mental events and that those prior events were themselves produced in the same manner by even earlier events etc. in an unbroken chain of cause and effect that can be traced back to before the person was even born. Determinism does not imply that people will not modify their behaviour in response to good reasons for doing so. It merely implies that whether a person recognises and responds to one particular reason for action rather than another at any given time is determined by prior events in the manner described above.

'Compatibilism' is the view that people can still have 'control' over their actions in the sense required for free will even if all their actions are determined by prior events over which they did not have control. Most compatibilists also believe that free will is compatible with indeterminism. 'Incompatibilism' is the view that free will is not compatible with determinism. 'Libertarian incompatibilists' argue that free will requires the falsity of determinism, that determinism is false and that people can be free.

Direct interventions and libertarian freedom

Most libertarians would oppose any type of direct intervention that guaranteed that the agent would act in one particular way. For most libertarians, freedom consists partly in the ability to choose between different alternatives for action, without the outcome of one's decision being guaranteed in advance by prior events.

In a very recent article, Harris (2011) has developed an argument against using direct neurological interventions to morally enhance offenders, which seems to rely on libertarianism. He begins by referring to the image of a forking path, stressing that if an agent is genuinely free then it is possible to choose to go down any one of the available paths. After quoting a passage from *Paradise Lost*, he, agrees with Milton that for agent to be capable of virtue it must be possible for him to do wrong or to 'fall'. He writes:

> Without the freedom to fall, good cannot be a choice; and freedom disappears and along with it virtue. There is no virtue in doing what you must...[Liberty could be] threatened by any measures that make the freedom to do immoral things impossible...sufficiency to stand is worthless, literally morally bankrupt, without freedom to fall...(Harris, 2010, 105-111).

Determinism entails that, given the facts of the past and the laws of nature, nobody could have acted differently from the way in which they actually did act. If a person actually refrained from doing an immoral thing, then given these facts and laws, it was physically impossible for them to have done the immoral thing. Harris's (2011) arguments seem to imply that causal determinism is incompatible with true virtue, because virtue requires that it was genuinely possible for the agent to have been vicious. There are, of course, compatibilist interpretations of alternative possibilities (which will be discussed below). However, Harris's (2011) arguments against direct interventions only seem to make sense on a libertarian interpretation. He states that it is a conceptual truth that God himself could not have *guaranteed* that human beings would behave virtuously and still have left us free (Harris, 2011). This is implied by libertarianism, which states that nothing, not even God, can ensure in advance that a free agent will decide to do one thing rather than

another – the future is open right up until the agent makes her choice. However, according to compatibilism, individuals who are predetermined to behave virtuously are still free (all that matters on this compatibilist view is that in some *hypothetical* scenario the agent would have behaved differently). It seems perfectly conceptually possible to imagine a world which God had created in such a way that *every* individual was guaranteed to develop into a virtuous agent and yet retain freedom in the compatibilist sense (see Pereboom, 2005 for a discussion of determinism and Christian theology).

There are various reasons why it is problematic to rely on libertarianism as one's only basis for opposing intuitively objectionable types of direct intervention. Firstly, Libertarian free will requires that certain empirical facts obtain. It requires that human deliberations are (at least sometimes) undetermined. It also requires that they are undetermined in a way that does not merely introduce randomness into our deliberations. Most libertarians themselves concede that we lack epistemic justification for these beliefs (see Double, 2002). Libertarianism therefore makes the question of whether anyone ever has free will a hostage to empirical fortune. For this reason, many philosophers and legal theorists do not wish to rely on the libertarian notion of free will.

A second difficulty with libertarianism is that it is far from obvious that the 'freedom to fall' is as crucially important as libertarians make it out to be. If someone does good things (e.g. helping others, telling the truth, speaking out against injustice etc.) because they genuinely recognise that there are good moral reasons for doing the right thing then it does not seem wholly inappropriate to call them 'virtuous', without enquiring into whether they were capable of doing morally obnoxious deeds. It may be that certain people have such a vivid awareness of the good (due perhaps to having received an inspiring moral education) that leading an immoral life is not a genuine psychological option for them. It does not seem obvious that such people necessarily lack a freedom that is really worth having.

Compatibilism part one: Free will and rational flexibility

As explained above, determinism does not rule out the possibility that agents will *modify their behaviour in the light of logically relevant data*. Flexibility – the ability to adapt one's behaviour in an appropriate way to changes in circumstances – is generally agreed to be a hallmark of rationality. Several compatibilist accounts of freedom emphasise the importance of this ability to respond to relevant reasons (which I will refer to as 'rational flexibility').

Certain types of direct intervention could undermine the offender's rational flexibility. Imagine, for instance, that the intervention instils an intense feeling of aversion to the idea of being violent towards others, which the agent cannot resist. The agent in this scenario lacks rational flexibility since there is no possible situation in which she would resist the aversion. The agent's behaviour is not sensitive to changes in her circumstances, nor to her other desires and beliefs which are relevant to her decision about how to act. She would not behave violently even if she believed that she had good reasons to do so (for example, if violence were the only way to save her own or someone else's life).

It might be objected that flexibility is not necessary for free will or rationality, because a person who is thoroughly committed to acting in accordance with a certain moral principle might not behave differently

under any circumstances. For instance, a committed pacifist might never resort to violence, and yet would be considered free and rational. However, it is possible to distinguish the offender with an irresistible aversion to violence from a person with a firm moral commitment. If a person's non-violent conduct is genuinely a response to a moral reason, then one of the causal factors bringing about her behaviour is the perception that violence is morally wrong. She has the capacity to alter her behaviour if she revised her moral position in the light of new arguments or evidence. This capacity to alter one's behaviour in response to a change in one's values is an important kind of flexibility.[68]

Various different accounts have been given of what possessing a 'capacity' would amount to in a deterministic world. Here is one version: An individual has a capacity to perform an action, if she possesses certain intrinsic properties (including properties of her brain) which would be (non-deviantly) causally operative in her performing the action if she chose (tried, decided or intended) to exercise this capacity and if the circumstances were favourable to the exercise of this capacity.[69]

This is a conditional account of capacity. It defines capacity in terms of what *would* happen *if* certain conditions obtained. Now, in some cases, the relevant conditions may never obtain. For instance, someone may be so firmly committed to certain values that there is no realistic scenario in which she would make the choice to abandon or flout them. Yet it may be that she would act in contravention of her values only *if* she chose to abandon or flout them. Nevertheless, many compatibilists would judge her free will to be intact in this situation, on the basis that she would behave differently if conditions were different, even though it is in fact impossible in the real world for the relevant conditions to be different.

This account fails to explain what is problematic about certain types of intuitively troubling intervention. For instance, consider an intervener who neurologically manipulates an individual's value system in a way that ensures that the manipulated individual will do exactly as the intervener wishes. Counterintuitively, the compatibilist account of freedom in terms of rational flexibility implies that this type of intervention would not necessarily violate the offender's free will. The person is free, on this account, as long as there are some *possible* reasons that would induce the person to behave differently. This condition would be satisfied even if

---

[68] Compatibilists differ over whether the flexibility possessed by rational agents in a deterministic world genuinely amounts to a capacity to behave differently from the way that one in fact behaves. The following theorists argue that it does: Fara, 2008; Vihvelin, 2004. The following theorists disagree, maintaining that the disposition to respond differently if different reasons were present is simply a feature of the way in which the agent *actually* behaves: Fischer and Ravizza, 1998.

[69] I am not suggesting that compatibilist accounts of 'capacity' are successful. This version in the text is based loosely on the account given by Vihvelin (2004). I have added the qualification that 'the circumstances must be favourable' in an attempt to take into account an objection raised by Clarke (2008).

the intervener had shaped the person's psychology in a way which ensures that the relevant reasons are ones that are unlikely *actually* to arise in the normal course of things. (For instance, the intervener might implant in the offender a belief that violence is morally wrong – a belief so firm that the offender would only resort to violence if her own or someone else's life were at stake.) This account of free will lacks the resources to account adequately for what is disturbing about this kind of manipulation. For this reason, some compatibilists supplement their accounts with an 'authenticity' requirement. Other compatibilists reject the flexibility model altogether in favour of an authenticity model.

Compatibilism part two: Freedom as authenticity

Authenticity and psychological coherence

The 'freedom as authenticity' approach defines 'free will' in terms of whether the agent's actions express her 'real self'. Compatibilists differ over which psychological states are to be identified with the agent's 'real self' (for a critique of Real Self Theories, see Wolf, 1990). Probably the most influential Real Self View was developed by Frankfurt (see e.g., Frankfurt, 1969). Frankfurt (1969) defined free will in terms of whether the agent's first order desires 'cohered' with the agent's second order desires. According to Frankfurt (1969), first order desires have actions as their objects. A person's desire to eat some chocolate is an example of a first order desire. Second order desires have first order desires as their objects. To say that a person has a second order desire to do something means that she wants to desire to do something. For instance, she might want to have the desire to exercise. According to Frankfurt (1969), in order to be free, an action must flow from a desire that the agent wants to have and which she wants to be executed in action. She must 'wholeheartedly identify' with the desire that results in her action. Other compatibilists, such as Watson (2004), focus instead on coherence between the agent's desires and values, rather than between different orders of desire.

Direct interventions could undermine an offender's psychological coherence. An intervention might cause the offender to have strong desires or aversions which jar with his values or second-order desires. For example, the intervention might cause the offender to experience powerful feelings of disgust at the idea of re-offending. The offender may not endorse or identify with these feelings of disgust. This kind of intervention creates an internal conflict between fundamental constituents of the person's agency – between his values and his desires/feelings. Alienation from his desires and feelings can threaten the person's identity, as it seems that an important part of his mental life is not truly his own.

It is important to remember, however, that direct interventions need not create psychological conflict within the offender. The offender may welcome the change in his motivations. In fact, a direct intervention might enhance an offender's psychological coherence, by bringing his feelings and desires more into line with his values. For instance, prior to intervention, the offender may have felt deeply ashamed of his violent impulses and may feel that interventions, which reduce the strength of those impulses, help him to become the sort of person he wants to be.

Furthermore, if we define 'free will' in terms of psychological coherence, then it seems that the following approach would preserve the offender's free will: employ direct interventions in order to modify *both* the offender's first-order desires *and* his second-order desires and values, in a way that ensures psychological harmony. Some philosophers, such Frankfurt (1969), accept this conclusion. Yet it would strike many people as counterintuitive to suggest that interfering to a *greater* extent in an individual's mental life and modifying aspects of the person that are *particularly central* to the individual's agency (i.e. their values) allows the individual *more* free will than interventions that only affect first-order desires/aversions. Some compatibilists have tried to avoid this counterintuitive conclusion by including a historical dimension in their theories.

Historical authenticity

According to historical compatibilists, whether a person's mental states are authentically hers at a given time depends on how she came to have those mental states. Her current mental states are only authentic, on this view, if they are connected in an appropriate way to the agent's earlier mental states. Thus, even if a direct intervention left the agent with a set of desires, beliefs and values etc. that were coherent and not in conflict, historical compatibilists might still find the intervention objectionable if the individual's post-intervention mental states were not appropriately connected to her prior mental states. What counts as an appropriate connection? At least three different types of connection have been suggested.

Similarity with previous mental states

Historical compatibilists often focus on cases where a significant alteration to the brain brings about a very sudden, dramatic change in the agent's motivations. Many different scenarios have been discussed, including: a very good woman who, after being manipulated by an evil neuroscientist, acquires the values of a serial killer (Mele, 2006) and a saintly nurse who, after receiving a blow to the head becomes cruel and reckless towards her patients (Tadros, 2005). They cite these examples as central cases where the individual's free will has been eliminated. There are also documented real-life examples of sudden personality changes, e.g. acquired paedophilia[70] (Burns & Swerdlow, 2003) and acquired sociopathy (Damasio, 1994).

Now, historical compatibilists acknowledge that sometimes ordinary people, whom we normally regard as possessing 'free will', undergo fundamental changes in their character, values, and desires. However, when such fundamental changes occur, they typically emerge gradually over time. Even if a person's

[70] I am thankful to an anonymous reviewer for suggesting this example.

motivational set-up when the agent is twenty years old differs considerably from her motivational-set up at fifty years old, this often is the result of a very gradual transformation where each incremental stage in the person's development resembles the previous stage in important respects, but where the final stage in the series is very different from the initial stage.

There are two problems with this version of historical compatibilism. Firstly, there are cases of individuals who undergo very fundamental changes in their values over quite a short period of time, and are still considered to be free. For instance, the individual may have a 'road to Damascus experience' – an inspired insight into important moral truths, which lead her to reject her previous values. This suggests that incremental change is not, in fact, a necessary condition for free will. Therefore, the fact that a direct intervention brings about a sudden change in the offender's values does not *in itself* render the offender unfree. Secondly it is possible to imagine a type of direct intervention that successfully alters the offender's values but which takes effect gradually over time. This version of historical compatibilism lacks the resources to explain why such an intervention is intuitively objectionable.

A connection in terms of deliberation

On this view, if an agent's values alter, the agent's new value is only authentic if the acquisition of this value was preceded by deliberation in the light of the person's prior value system (Haji & Cuypers, 2007). However, road to Damascus cases provide a challenge for this view as well. Imagine that an agent, Denise, was a thoroughly selfish person with a corrupt value-system. One day a natural disaster strikes her town. She is unharmed but encounters numerous victims of the disaster. Denise experiences an unfamiliar experience of compassion accompanied by a sudden insight into the reasons for helping others. She acts on her new moral insight and performs some good deeds. However, she did not deliberate about her new insight in the 'light' of her old corrupt value-system. The new moral insight just displaced the old corrupt values. Is Denise's insight therefore inauthentic and are her subsequent actions unfree? It does not look that way.

Imagine Denise's community decides to present her with a medal for her good deeds. At the awards ceremony, a psychologist stands up and says, 'As part of my research into why people perform heroic acts, I have looked very carefully into Denise's case. I discovered that when Denise acquired her new, emotionally-charged awareness of the need to alleviate human suffering, she did not evaluate this insight in the light of her earlier corrupt value-system. In fact, her corrupt evaluative scheme was completely idle! Hence her new good moral values are inauthentic and the actions that flowed from them were not an exercise of free will. Denise therefore does not deserve a medal.' This reaction would seem bizarre. Therefore, the 'deliberation connection' does not seem to be a necessary condition for free will and the supposed absence of this connection *per* se cannot provide a convincing basis for objecting to direct neurological interventions.

Mental states connected in virtue of sharing the 'same kind of mechanism'

According to Fischer and Ravizza's (1998) version of historical compatibilism, in order for an agent's actions to be genuinely her own, the agent must have previously 'taken responsibility' for the mechanisms

from which her actions arise, by *viewing herself* as being responsible for actions that flow from these mechanisms. By 'mechanisms', they mean the features of her agency that play a causal role in her actions (including, but not limited to mental states such as intentions, desires and beliefs). On Fischer and Ravizza's (1998) view, when an agent, at a particular time, comes to take responsibility for behaviour that flows from a certain type of mechanism, she *thereby* takes responsibility for her future behaviour that results from the *same* kind of mechanism. They claim that motivations resulting from direct neurological interventions (almost invariably) involve a *different kind of mechanism* from ordinary motivations. Therefore, they maintain, when an individual takes responsibility for her ordinary mechanisms she does *not* thereby typically take responsibility for motivations or actions that arise from neurological interventions.[71]

The problem arises when Fischer and Ravizza (1998) try to explain what makes a mechanism belong to one 'kind' rather than another. They do not simply maintain that actions which flow from psychological states like 'desires', 'beliefs' and 'intentions' arise from one type of mechanism and actions that have nothing to do with such psychological states (such as epileptic seizures) belong in a different category. If they settled for this simple account then it would not help them to differentiate reliably between cases of 'ordinary' mechanisms and mechanisms produced by intuitively objectionable types of direct interventions. For it is possible to imagine mental states such as desires and beliefs being induced by direct interventions.

Fischer and Ravizza (1998) rely heavily on intuition to differentiate between different kinds of mechanism. They maintain that, *intuitively*, motivations resulting from direct stimulation of the brain belong (in most cases) to a different kind of mechanism from motivations that are determined in the 'ordinary' way by one's genes and environment. This approach is open to challenge. For it seems that the notion of 'different mechanisms' is no longer doing the work it was supposed to do. This notion was meant to help *explain* why we intuitively feel that certain types of direct interventions are problematic. But instead it seems like our intuitions that certain types of direct interventions are problematic dictate whether one mechanism counts as belonging to a 'different kind' of mechanism from another. In order for the notion of 'different mechanisms' to have explanatory power, Fischer and Ravizza need to have a principled basis for individuating mechanisms, which is derived from *"independent reflection on the nature of these mechanisms"* (Pereboom, 2006, 200; see also McKenna, 2001). Otherwise, it seems that they are merely stipulating that certain mechanisms are different from others in an *ad hoc* way in order to generate the conclusions they want about direct interventions. Unfortunately, it is far from obvious that truly independent criteria for individuating mechanisms (e.g. derived from psychology, or neurology) will produce the results that Fischer and Ravizza desire.

[71] Fischer and Ravizza (1998) do acknowledge that in certain cases agents can take responsibility for their post-manipulation mechanisms. I am thankful to an anonymous reviewer for reminding me of this point.

So far, this paper has examined various free-will-based objections to direct interventions and has identified certain difficulties with all of them. This should provide some motivation for those opposed to direct interventions to look for a way of objecting to direct interventions that avoids the problematic issues surrounding the notion of free will. The following section of this paper outlines such an objection.

Reform, free will and moral status

The thought experiment

Consider the following thought experiment: One day an angel appears on earth. The angel possesses a magic flute. Anyone who hears the flute will suddenly have a powerful insight into fundamental moral truths. This vivid recognition of the reasons for behaving morally will motivate the agent to act in accordance with these reasons. Flute in hand, the angel marches off to the nearest prison. The authorities get to hear about this before the angel reaches the prison. What should they do? It seems that the free-will-based objections to direct interventions would apply equally to the magic flute scenario – if the recognition of moral reasons and the subsequent commitment to act accordingly, *guarantees* that the offender will act virtuously (in the actual world) then this violates incompatibilist freedom; if the offender's new values are disconnected from her prior values (in any of the senses of 'disconnection' mention above) then this violates a version of 'freedom as authenticity'; given that a causal factor behind the change of values (listening to the flute music) does not provide the agent with any new reason for changing her behaviour, then this arguably goes against a rationality-based conception of free will. If these approaches to free will are correct, then it seems that the authorities have great cause for concern - the free will of a large number of offenders is in jeopardy. Yet it seems counter-intuitive to suggest that the authorities would have a pressing obligation to rush to prevent the offenders from being affected by the music's reformative powers, or that it would be such a terrible thing if the authorities failed to take action in time to prevent the prisoners from being reformed.

The 'magic flute' thought experiment is intended to cast doubt on the claim that changing an offender's values using direct interventions, rather than moral dialogue, necessarily violates the offender's free will in an objectionable way. This thought experiment features a means of altering values that does not involve moral dialogue and yet does not seem to violate the offenders' free will, or even if it does so, it does not seem seriously morally objectionable. However, this thought experiment does not show that it is all right for *us* to use interventions other than moral dialogue. It is submitted that ordinary human beings do not have the *moral status* to directly re-shape a person's values or goals using means other than rational persuasion. The objection to direct interventions presented in this paper does not rely on the idea that these interventions violate the offender's 'free will', conceived of as a capacity that we can identify just by examining the individual's psychology and actions carefully enough. Rather, it is submitted that an objection to such interventions can be based on the problematic nature of the *relationship* between the intervener and the subject of the intervention.

It is possible to identify the objectionable features of this relationship by highlighting the ways in which it departs from a model of an appropriate type of relationship between the state and offenders. I will not

attempt to fully describe and defend such a model within the scope of this paper. Rather, I will present certain principles concerning how the state ought to relate to offenders, which have some intuitive plausibility. If my account is accepted, it provides a basis for objecting to certain kinds of direct neurological intervention, which does not rely on the notion that these interventions violate the offender's free will.

It is submitted that society's response to criminal behaviour should recognise that offenders are members of the moral community, albeit members who have breached the community's norms. This principle is supported by the intuition that the state should not 'objectify' law-breakers – that offenders should be treated as persons and not as things. Objectifying a group of people can involve emphasising that 'they' are fundamentally unlike 'us'. It can involve focussing on the idea that a deep division exists between the objectified group and the rest of society. One way of respecting offenders' personhood is to highlight commonalities that still exist between the offender and the other members of the community, and to preserve certain connections between the offender and other moral agents. This kind of respect can be shown through engaging rationally with the offender as he is, and by challenging his mistaken views with arguments, without using direct neurological interventions to fundamentally re-shape his psychology. There are several ways in which rational dialogue affirms commonalities between offenders and other moral agents.

Dialogue and equality

Engaging in dialogue with the offender includes him within the moral community by allowing the offender to voice his criticisms of the community's norms, which can potentially contribute to a shift in those norms. Dialogue leaves open the possibility that *either* party may change the other. As Stern (1974) writes:

> [Dialogue] involves the recognition of a certain equality between oneself and the other. There is, in general, no point in reasoning unless the other person is capable of seeing reason, getting the point. If he can do that, he can also correct *me* if I am mistaken (Stern, 1974, p.75).

In contrast, attempting to re-shape the offender's values using direct neurological interventions is a one-way street. It seeks only to change the offender, to ensure that he will think and act in a particular way.

The most appropriate way for members of a moral community to attempt to change one another is through dialogue. Engaging offenders in dialogue, rather than re-shaping offenders' values through direct interventions, assumes that there is a commonality between the offender and other moral agents. It implicitly acknowledges that the authorities (and majority opinion) are fallible, as is the offender. It allows that the offender (as he is, without neurological modification) may have useful insights, as other agents do. It also allows for the fact that the pursuit of moral understanding is a shared process. People need to interact with other people and to consider different points of view before they can form reliable judgements about how they should act.

The above considerations do not apply to the case of the angel in the thought experiment. The angel, as the embodiment of rationality and virtue, never stands in need of 'correction'. In contrast, the authorities do not have the moral status to portray themselves as the embodiment of rationality and virtue. Re-shaping offenders' values through direct neurological interventions replaces the acknowledgement that the authorities

(like the offender) are human and fallible with the inappropriate assumption that the authorities are absolutely certain about what the 'right' values are. Furthermore, the angel is not a fellow member of the offender's human community, so the lack of dialogue between the angel and the offender does not convey the message that the offender is excluded from the community. However, if other human beings were to re-shape offenders' values through direct neurological interventions, rather than engaging them in dialogue, this would be an act of excluding offenders from the moral community.

Focussing on the principles that should govern the *relationship* between the state and the offender helps to explain why the use of direct interventions by the state is more intuitively troubling than the intervention employed by the angel in the thought experiment. The relationship-based approach also produces other intuitively appealing results. Unlike some of the free-will-based approaches discussed above, the relationship-based approach implies that more extensive modifications of the offender's motivations are worse than less extensive interventions. For instance, interventions that just enhance the offender's control over his violent impulses, or reduce the strength of those impulses do not alter the offender's values, and so leave open the possibility that he will criticise the authorities on the basis of those values.[72] Modifying the offender's values precludes this possibility. Modifying the offender's values also sends out the strong message that the authorities view themselves as having hugely privileged access to knowledge of what the 'right' values are.

It might be objected that this paper takes an unrealistic view of the potential for offenders to make a valuable contribution through moral dialogue to society's understanding of moral norms. Surely the authorities can be very confident that some offenders are completely in the wrong and that some of society's norms are very well-founded. In response, it is important to remember that, historically, a number of values which society has now come to reject once seemed self-evidently sound and that individuals who were very widely condemned by the rest of society have ultimately been vindicated.

Furthermore, instituting a policy of trying to instil acceptable moral values in offenders through direct neurological modification would create a disturbing relationship between the state and offenders, even if the policy were restricted to offenders who were genuinely in the wrong, and even if it succeeded in instilling values that were genuinely well-founded. Such a policy would mark a huge shift towards characterising these offenders as 'the other' and thus towards objectifying them. It would express the attitude that they are a group of people to whom we need not listen (or at least that we need not listen to them until we have modified their brains such that they are likely to tell us what we want to hear). If all attempts to change offender's values involve entering into a relationship with the offender, rather than relying on direct

[72] Although dialogue has an advantage over even this technique, in that dialogue, unlike direct interventions, positively reaffirms the offender's status as a moral agent and includes him within the moral community.

neurological interventions, then society is less likely to lose sight of the personhood of the offender. In addition, even if society's condemnation of a particular offender is justified and the offender is completely in the wrong, dialogue with the offender can still make a useful contribution to other agents' moral understanding. For the attempt, through rational dialogue, to reform a wrongdoer who is very unwilling to be persuaded can cause the would-be reformer to try to make his arguments as compelling as possible, which can lead to a clearer understanding of the justification for society's norms.

It might also be objected that this paper adopts an excessively rosy view of the available alternatives to direct interventions. No society responds to criminal behaviour by relying on dialogue alone. A prison sentence, for instance, *"is more than an appeal to sweet reason and morality"* (Stern, 1974, p. 82). Furthermore, it might be argued, punishing criminals necessarily involves highlighting the differences (rather than commonalities) between offenders and law-abiding citizens, by condemning the offender as a wrongdoer. Punishment also excludes offenders from the community. It can do this in terms of the moral stigma that attaches to a criminal conviction and sentence. It can also physically exclude the offender from the community, e.g. by putting him in prison.

It should be acknowledged that society's response to criminal behaviour does involve coercion, exclusion and the highlighting of differences between offenders and law-abiding citizens. It is perhaps impossible to conceive of a practicable approach to the problem of crime which does not involve these elements to some degree. But it is submitted that society's response to criminal behaviour can and should *also* involve dialogue (and not just coercion); that it should emphasise the commonalities between offender and other moral agents (and not just the differences); and that it should preserve some connections between the offender and the rest of the community (and not exclude the offender entirely).

It is important to emphasise that this paper is not a defence of our current system of punishment and rehabilitation. Some of our current approaches to dealing with criminal behaviour are objectionable and fail to treat the offender as a member of the moral community. In order for our practices to be justifiable they would have to include much more sustained attempts to engage with offenders, to present them with moral reasons for changing their behaviour and to re-integrate them into the community.[73]

Nevertheless, the employing of coercion as part of society's response to criminal behaviour is *compatible* with continuing to view the offender as a member of the moral community, by among other things, allowing the offender to challenge the authorities, on the basis of his pre-existing value-system. In contrast, as argued above, the technique of re-shaping the neurological basis for offenders' values would take a significant step towards characterising the offender as 'the other'. It would vastly increase the (already

[73] For some criticism of the current system and for one account of ways in which it should be reformed which emphasises the importance of moral dialogue with offenders see Duff (2001).

considerable) powers for controlling offenders' behaviour which the authorities have at their disposal. This would set the authorities on a completely different plane from offenders, greatly increasing the inequality of power between them.

Dialogue and offenders' better natures

There is a further way in which including moral dialogue in society's response to offenders, rather than employing direct interventions, emphasises the commonalities between the offender and the rest of the community. Dialogue aimed at persuading offenders to reform typically involves appealing to the offender's 'better nature'. This presupposes that, in common with most law-abiding citizens, the offender has certain positive qualities and that, although he committed a serious wrong, he is not completely corrupt.[74] In contrast, direct Interventions imply that offenders are different from law-abiding individuals in a very fundamental way. They imply that offenders are so inferior to the rest of the community in terms of their moral characters that these offenders will not respond appropriately to the most compelling moral reasons for changing their behaviour (unless the offenders receive radical neurological modifications). Most moral agents assume that, even though they may have certain vices and may sometimes behave wrongly, they would respond to really compelling reasons for improving their behaviour, provided that they were given sufficient time to reflect on the matter, that the reasons were put to them persuasively enough and the issue at stake was really important. They further assume that responding in this way is possible for them because they are not thoroughly bad; that they respond to these compelling moral reasons because they already have certain good qualities, which are brought out by sufficiently persuasive arguments. Viewing oneself in this way is particularly valuable, because it provides an important basis for self-respect. The preparedness to re-shape offenders' values through direct neurological interventions suggests that offenders lack the qualities that provide this basis for self-respect.

It should be noted that possessing these positive moral qualities is not the same thing as 'having free will'. It is conceivable that an individual might improve his behaviour of his own free will, even if hitherto he had been thoroughly corrupt. It is not essential to the common sense notion of free will that a person's moral improvement was partly caused by the fact that the individual already had certain good moral qualities. But, as a matter of fact, most instances of moral improvement probably do build on pre-existing good qualities and it is part of a positive self-conception to view one's moral development in this way. Extensively re-shaping an offender's values through direct neurological interventions strongly suggests that the authorities consider the offender's existing character to be so comprehensively morally inadequate that positive moral change is unlikely to emerge from it. This carries the message that offenders are fundamentally not like 'us'.

[74] A similar point is made in Duff (1986, p. 266).

This message is much more extreme than the alternative message (conveyed by moral dialogue) that the offender behaved wrongly on a particular occasion, or that he demonstrated a particular vice.

A critic of my view might raise the following objection. My argument stresses that membership of the moral community is valuable. It also accepts that, in some cases it seems fairly likely that the offender's capacities for practical reasoning are such that they will never lead the offender to be reformed and to be genuinely restored to the moral community. Yet, if this is the case, then it would surely benefit such an offender if it were possible to use direct neurological interventions to re-shape his psychology such that he is much more likely to fully grasp and take to heart the moral reasons for reforming. Would not increasing the probability that the offender will *actually* be restored to the moral community in this way be better for the offender than maintaining the *fiction* that it is possible that he will reform, when in fact it seems that he never will?

In response, while it may be true that the individual offender might benefit from this kind of intervention, the objection to direct interventions that this paper advances is not based primarily on the idea that direct interventions violate the individual's rights or interests in every case. Rather, this paper maintains that the use of certain types of direct intervention would create a troubling relationship between different groups within society. A policy of employing direct interventions to re-shape offenders' values would be based on the assumption that these offenders' existing capacities for moral agency are so fundamentally inferior to the capacities of the rest of the moral community that these offenders will not respond appropriately to the most compelling reasons for changing their behaviour. Basing social practices and institutions on the assumption that a particular group of individuals are radically incomplete as moral agents goes against the ideal that the moral community should be as inclusive as possible and that it should emphasise its members' common humanity. Incorporating into our social structures the message that a particular group is so different from the rest of us that they require radical neurological modification to enable them to be part of the moral community is *prima facie* objectionable even if such a system would end up (in a sense) benefitting certain offenders.

For the reasons stated above, it is also submitted that altering an offender's values using direct neurological interventions would be unacceptable even if the offender requested such treatment. The offender's consent could not legitimise this practice because the practice affects society's stance towards offenders as a group. The very act of offering this type of intervention to offenders would send out the message that all offenders who are offered the intervention stand in need of it, whether or not they ultimately agree to it. This practice has the potential to be socially divisive and its effects are not limited to those offenders who give their consent. Therefore the offender's consent is not sufficient to make it morally acceptable.

Different types of intervention

Finally, it should be reiterated that this paper does not oppose the use of *all* neurological interventions within the criminal justice system. It specifically focuses on the idea of using neurotechnologies to deal with offenders who are basically rational, in an attempt to directly re-shape these offenders' values or goals.

There are other possible types of neurological interventions which are less vulnerable to the challenges raised in this paper (although they may face different challenges). For instance, neurotechnologies might be used in order to treat offenders who have mental illnesses. Or neurotechnologies might be used to enhance the rational capacities of basically normal offenders so that they are better able to decide for themselves which goals they should pursue, or which values they should endorse.

Having said this, there is not always a razor sharp line between using neurotechnologies to directly re-shape offenders' values/goals (which I oppose) and the use of these techniques to enhance offenders' rational capacities (which are arguably less problematic). For example, an intervention might reduce the strength of an offender's violent and/or deviant sexual impulses. It might be argued that this is a method of enhancing offenders' rational capacities, because intense, repetitive urges or fantasies can cloud an individual's judgement, making practical reasoning difficult. Reducing the strength and frequency of these urges could put the offender in a better position to focus on the reasons that are relevant to his decision about how he should act. Alternatively, it might be argued that interfering with offenders' urges is a method of directly re-shaping their values, because an offender who values violence or deviant sexual conduct might do so partly as a result of experiencing these impulses and urges. It is impossible within the scope of this paper to give a detailed account of how these borderline cases should be dealt with. However, it should be noted that this issue can be decided partly on the basis of the principles that have already been outlined in this paper. One relevant consideration is the amount of control which the intervention would allow the state to exert over the agent's decisions about what he should do. The greater the state's level of control, the greater the inequality between the offender and the rest of the community. Interventions which merely reduce the strength of an offender's violent impulses, do not give the state the power to ensure that the offender endorses the state's favoured values. The offender may still reject society's demands. Such interventions are therefore less troubling than interventions that allow the state to shape the offender's behaviour and inner life to a greater extent. Furthermore, interventions that would alter an attribute which is *central* to who the person is, as an agent, are particularly troubling. A particularly fundamental alteration sends out a strong message that the offender is radically defective, and unlike the rest of 'us'. Again, it is submitted that a momentary impulse, or urge is less central to the offender's agency than, say, a second-order desire, or a firm commitment to a particular principle or course of action.

This paper has focussed on explaining why certain extreme types of neurological intervention are *unacceptable*. The question of whether *any* form of neurological intervention within the criminal justice system is acceptable, all things considered, is a complex issue. A full discussion of this topic is outwith the scope of this paper (for a more detailed discussion see Shaw, 2011).

Conclusion

This paper has attempted to expose some serious difficulties faced by 'free-will-based' objections to attempting to 'morally enhance' offenders using direct neurological interventions. It suggests a different basis for objecting to extreme interventions that aim to directly alter offenders' values and basic goals. The

approach advocated in this paper emphasises the value of moral dialogue and of treating the offender as a member of the moral community.

References

Burns, J.M., & Swerdlow, R.H. (2003). Right orbitofrontal tumor with pedophilia symptom and constructional apraxia sign. *Archives of Neurology, 60,* 437-440.

Clarke, R. (2008). Dispositions, abilities to act, and free will: The new dispositionalism. *Mind, 118*, 323-351.

Damasio, A. (1994). *Descarte's Error: Emotion, reason and the human brain.* New York, NY: Putnam.

Double, R. (2002). The moral hardness of libertarians. *Philo, 5*, 226-234.

Douglas, T. (2008). Moral enhancement. *Journal of Applied Philosophy, 25*, 228-245.

Duff, R. A. (1986). *Trials and punishments*. Cambridge: Cambridge University Press.

Duff, R. A. (2001). *Punishment, communication and community.* Oxford: Oxford University Press.

Fara, M. (2008). Masked abilities and compatibilism. *Mind, 117*, 843-865.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility.* Cambridge: Cambridge University Press.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy, 66,* 829-839.

Haji, I., & Cuypers, S. (2007). Magical agents, global induction, and the internalism/externalism debate. *Australasian Journal of Philosophy, 85,* 343-371.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics, 25*, 102-111.

Harrison, G. (2010). A challenge for soft line replies to manipulation cases. *Philosophia, 38*, 555-568.

Levy, N. (2007). *Neuroethics: Challenges for the 21st century*. Cambridge: Cambridge University Press.

McKenna, M. (2001). Book review: Responsibility and control: A theory of moral responsibility, by John Martin Fischer and Mark Ravizza. *Journal of Philosophy, 98,* 93-100.

Mele, A. (2006). *Free will and luck.* New York, NY: Oxford University Press.

Moore, M. (1985). Causation and the excuses. *California Law Review, 73*, 1091-1150.

Pereboom, D. (2006) Reasons-responsiveness, alternative possibilities and manipulation arguments against compatibilism: Reflections on John Martin Fischer's *My Way. Philosophical Books, 47,* 198-212.

Pereboom, D. (2005). Free will, evil, and divine providence. In A. Chignell & A. Dole (Eds.), *God and the ethics of belief: New essays in philosophy of religion* (77-98). Cambridge: Cambridge University Press.

Stern, L. (1974). Freedom, blame, and moral community. *The Journal of Philosophy, 71*, 72-84.

Shaw, E. (2011). Cognitive enhancement and criminal behaviour. *Proceedings from Cognitive enhancement: An international conference for young scholars.*

Tadros, V. (2005). *Criminal responsibility.* Oxford: Oxford University Press.

Vihvelin, K. (2004). Free will demystified: A dispositional account. *Philosophical Topics, 32,* 427-450.

Watson, G. (2004). Responsibility and the limits of evil: Variations on a Strawsonian theme. In G. Watson (Ed.), *Agency and answerability: Selected essays* (219-259). New York, NY: Oxford University Press.

Wolf, S. (1990). *Freedom within reason*. New York: Oxford University Press.

Chapter 9
Cheating with implants: Implications of the hidden information advantage of bionic ears and eyes

Bert-Jaap Koops
Tilburg University
Tilburg Institute for Law, Technology, and Society
✉ e.j.koops@uvt.nl

Ronald Leenes
Tilburg University
Tilburg Institute for Law, Technology, and Society
✉ r.e.leenes@uvt.nl

Abstract Medical technology advances rapidly. As of 2009, about 188.000 people worldwide had received cochlear implants, and promising trials have been conducted with retinal and subretinal implants. These devices are meant to (partially) repair deaf and blind people's impairments, allowing them to (re)gain 'normal' sensory perception. These medical devices are ICT-based and consist of a sensor that transforms sensory data (auditory, visual, tactile) into signals that can be processed by the brain. Besides data from the regular sensors, in principle, also other data from other sources can be channelled to the brain through the implant, for example wireless data input from distant locations or even the Internet to prompt the bearer with instructions or information. This can be done without others present being aware of this form of techno-prompting, which might give the bionic person a competitive advantage in for instance meetings or negotiations. The medical implants could therefore be used for non-medical purposes somewhere in the future. This paper discusses the normative implications of this hypothetical form of human enhancement, focusing on aspects that are particularly relevant to this type of enhancement as compared to existing and well-discussed other forms of enhancement. In particular, we discuss information asymmetries, ethical aspects related to human enhancement, and some legal issues where the information advantage of bionic sensory implants could make a difference. Based on this discussion, we highlight questions for further reflection and provide some suggestions for the regulatory response to address the challenges posed by the future of bionic sensory implants.
Keywords human implants, neural prosthetics, information asymmetry, human enhancement

Introduction

Picture a group of security guards surrounding heads of state with dark glasses and curly black wires connecting their earpiece to the radio in their pocket. These men in black partially receive their orders from intelligence officers in control rooms that bring together the various intelligence sources surrounding the event. We are all aware of the fact that the men in black have prompters, and perhaps part of the effectiveness of this equipment stems from its visibility, similar to the disciplining effect resulting from the knowledge of being observed (Foucault, 1978). But in any case, the earpieces give their users an information advantage.

Of course anyone can buy such earpieces and be in constant contact with helpers outside the scene. This would potentially provide them with a similar informational advantage over others present.

This kind of human enhancement would possibly be frowned upon by others and potentially even be disapproved of. But what if people could have prompters without others being aware of this kind of enhancement?

Medical technology advances rapidly. As of 2009, about 188,000 people worldwide had received cochlear implants, and promising trials have been conducted with retinal and subretinal implants. These implants are ICT-based and consist of a sensor that transforms sensory data (auditory, visual, tactile) into signals that can be processed by the brain. They are meant to (partially) repair deaf and blind people's impairments, allowing them to lead a more 'normal' life. The use of these *"neural prosthetics"* (Merkel et al., 2007, p. 485) could in the future go beyond achieving 'normalcy'.

Instead of connecting regular sensors (microphone and camera) to the implant to create or restore the signal path between the external world and the brain, also other input can be used. Neural prosthetics can *"not only restore severed sensory functions, but also computer-enhance human capabilities"* (Merkel et al.,

2007, p. 143). For instance, audio streams from distant locations can be directly fed into the cochlea prompting the bearer with instructions or advice, similar to the instructions given to the security guards or presenters on television. The (sub)retinal implant can function as an internal 'head-up display' for visual data provided by remote sources. Since the data in these cases is fed directly to the brain, this can be done without others present being aware of this form of techno-prompting.

The implications of cochlear and (sub)retinal implants and their potential of being connected wirelessly to outside data sources have, as far as we are aware, not yet been discussed in the literature. It is important to analyse these implications with due attention to the specificity of the problems raised by this particular type of human enhancement, for, as Bostrom and Savulescu (2009) note, we need a 'contextualised and particularised' approach to addressing the question how to deal with human enhancement, applying a case-by-case analysis.

In this paper, we will therefore focus on the question: what are the normative implications of neural prosthetics with their potential for knowledge enhancement? This is a hypothetical discussion, since the technology does not yet exist to connect cochlear or retinal implants wirelessly with outside data sources. It is nevertheless important to start such a hypothetical and perhaps seemingly unrealistic discussion early on, since *"[i]f there is a single lesson to be learned about the past century of scientific and technological discovery, it may well be that the unimaginable rapidly becomes the commonplace"* (Garland, 2004, p. 29).

This article provides an explorative account of normative implications of neural prosthetics without aiming to be systematic or exhaustive. We approach the topic from the angle of ethics and law, indicating general issues, primarily from the outlook of liberal constitutional democracies.

This paper is structured as follows. We start with a discussion of implant technology state-of-the-art and use some cases. The core of the paper then discusses various normative implications of bionic sensory implants, focusing on those aspects that seem to be particularly relevant to this type of enhancement as compared to existing and well-discussed other forms of enhancement. We

distinguish between issues related to information asymmetries, other ethical issues related to human enhancement, and some legal issues. We conclude with pointing out issues that require further reflection in the academic and societal debate about sensory implants, and provide some suggestions that may help address the regulatory challenges posed by the future of bionic sensory implants.

Implant technology

The idea of brain implants already has a considerable history. Merkel et al. (2007, p. 120) quote José Delgado and his colleagues, who in 1976 remarked that with:

> …the increasing sophistication and miniaturization of electronics, it may be possible to compress the necessary circuitry for a small computer into a chip that is implantable subcutaneously. In this way, a new self-contained instrument could be devised, capable of receiving, analysing, and sending back information to the brain (Delgado et al., 1976 in Merkel et al., 2007, p. 120).

Today, these implants are a reality, although still in their infancy. The *"most advanced central neural prostheses today comprise the auditory implant, the visual implant, and the human-computer interface (HCI)"* (Merkel et al., 2007, p. 121). Cochlear implants (or bionic ears) are the most mature of these three types. The history of cochlear implants goes back to the late 1950s when André Djourno and Charles Eyriès placed wires on nerves exposed during an operation.[75] In 1972, a single-electrode implant designed by Dr. House and 3M was the first to be approved for implantation into adults by the US Food and Drug Administration (FDA). Later implants use up to 23 electrodes.[76] Getting a cochlear implant and learning to hear with it is slightly more challenging than accommodating to Vogon speech by slipping a Babel-fish – a tiny fish that translates all languages into your own – into one's ear (Adams, 1980). A cochlear implant generally consists of one or more microphones, a speech processor that filters and processes the sounds into signals that can be transmitted through a coil held in place by a magnet behind the external ear to a receiver and stimulator secured in the bone beneath the skin. From there the signals are sent to a spiral of electrodes threaded into the cochlea to stimulate the auditory nerve. The patient will have to learn how to interpret the implant's signals – they have to learn how to hear. The quality of hearing with a cochlear implant is (much) lower than hearing people experience. This is not surprising given that

[75] http://en.wikipedia.org/wiki/Cochlear_implant

[76] The Cochlear Nucleus CI500 uses 22 contacts allowing for detecting 161 different frequencies (see http://www.cochlear.com/uk/nucleus-cochlear-implants-0 for details). The Advanced Bionics HiRes 90K uses 16 contacts (see http://www.advancedbionics.com/CMS/Products/HiRes-90K/ for details).

even modern cochlear implants have at most 24 electrodes to replace the 16,000 hair cells that are used for normal hearing. However, the sound quality delivered by a cochlear implant is often good enough that many users do not have to rely on lip reading.[77]

Visual implants (or bionic eye), or visual prosthetics, work in a similar way. They consist of an external (or implantable) imaging system (camera), which acquires and processes visual information. The processed data is then transmitted to the implant wirelessly by the external unit (along with power for the implant). The implant converts the digital data to an analog output, which is used to electrically stimulate the visual system. The stimulation can be done anywhere along the optic signal's pathway, hence from retina via optical nerve to visual cortex. A major step was achieved when Giles Brindley (Brindley & Lewin, 1968) implanted an 80-electrode device on the visual cortical surface of a 52-year-old blind woman, allowing her to see 'light' (phosphenes) in 40 locations in her visual field. In 2010, the company Retina Implant reported success with a sub-retinal implant consisting of a 1500-electrode array. One of their patients, a 45-year-old Finland-based male, reported: 'As I got used to the implant, my vision improved dramatically. I was able to form letters into words, even correcting the spelling of my name. I recognized foreign objects such as a banana and could distinguish between a fork, knife and spoon. Most impressively, I could recognize the outlines of people and differentiate heights and arm movements from 20 feet away.'[78] As with cochlear implants, visual implants have a significant difference in quality of vision compared with normal eyesight, due to the enormous difference in number of receptors, the human retina consisting of an estimated 125 million receptors.

The third type of prosthetic implants are sensory/motor prosthetics. Electrode arrays can be implanted into (median) nerves. The array can be used to pick up signals from the underlying nerves, as well as to pass signals to these nerves. In 2002, scientist Kevin Warwick (Warwick et al., 2003) had an array of 100 electrodes implanted in his arm that allowed him to have a robot arm mimic the actions of his own arm, as well as have his arm respond to external signals.

Finally, and more futuristically, there are direct neural interfaces or brain/computer interfaces (BCI). BCI research is aimed at connecting the nervous system directly to computer systems rather than to devices such as cameras and microphones. Until now, BCI research has focused on recording signals from and providing stimuli to animal brains (rats, cats, and monkeys). For instance, Stanley, Li and Dan (1999) have shown to be able to generate movies of what their cats saw and to reconstruct recognisable scenes and moving objects on the basis of signals from their visual cortex. Wessberg

---

[77] http://en.wikipedia.org/wiki/Cochlear_implant

[78] http://www.businesswire.com/portal/site/home/permalink/?ndmViewId=news_view&newsId=20100317005294&news Lang=en.

and colleagues (2000) have developed BCIs that decoded brain activity in owl monkeys and used the devices to reproduce monkey movements in robotic arms. Moving from there,

> though futuristic, downloading of the brain may be also relevant in this context. With the possibility of downloading through modified techniques such as those described by Nicolelis at Duke, information from specific areas of the brain, or the whole brain, can be downloaded into a computer. Once downloaded, the information may be modified – for example, by adding a language capacity. Then the altered material may be uploaded into the same individual's brain...This return downloading may in time prove to be an excellent way to enhance cognitive skills (Tancredi, 2004, p. 102).

All these implants start out with the prospect of having people regain or gain capabilities of 'normal' people. As will be clear from the discussion, the implants involve sophisticated signal processing equipment to transform external signals (visual, auditory, sensory) to stimuli for the nervous system. Provided that the number of electrodes in the various types of implants increase significantly, roads are opened for other applications. Information can be superimposed onto visual information provided by the camera to create augmented reality-like applications directly on the retina or visual cortex. Think, for instance, of turn-by-turn information provided by a navigation system projected on one's visual system, or meta-information about the object being sighted that is directly fed into the brain along with the visual information. The path from medical implants to human enhancement, then, is relatively smooth: any information source that can be transcoded into signals suitable for the electrodes connected to the nervous system, can be employed for improving sensory perception.

Ethical and legal issues

This section discusses ethical and legal issues of neural sensory prosthetics.[79] We will briefly touch upon general issues that arise in other contexts as well, notably with implants and brain enhancement, but we shall particularly focus on issues that might play out differently, or acquire additional salience, when it comes to sensory implants being fed wirelessly from outside data sources. A key difference with other types of brain enhancement that have been discussed in the literature, is that sensory implants enhance by feeding content into the brain, not by enhancing the brain's capacity for processing content. This raises partly different types of question than those triggered by, for example, enhancing brain functionality through psychopharmaceuticals or brain stimulation. First and

79 We will use the terms neural sensory prosthetics and bionic sensory implants interchangeably.

foremost, ethical questions relating to information asymmetries arise. Second, several issues, primarily ethical, relate to the enhancement aspect of bionic implants. Finally, legal issues are discussed relatively briefly, since most of these are not specific to the implants under discussion.

Ethical issues relating to information asymmetry

Human decision-makers lack the ability and resources to arrive at optimal solutions to problems they are facing. Instead, due to cognitive and time constraints, they necessarily have to simplify the choices available and arrive at satisfactory rather than optimal solutions (Simon, 1947). Bounded rationality affects any problem-solving or decision-making. In interactions with other people, for instance in negotiations, additional factors contribute to not reaching satisfactory results. For instance, although people negotiate all the time, most are not trained to successfully do so (Thompson, 2005). Hence, they have difficulty in framing, structuring, and thinking creatively about solutions. There are also structural barriers, such as negative emotions and bad atmosphere and power imbalances (Mastenbroek, 1999; Moore 2003). People furthermore are hampered by cognitive barriers, such as loss aversion (Tversky & Kahnemann, 1981), the anchoring effect and overoptimistic overconfidence (Neale & Bazerman, 1991).

Some of these issues can be overcome by providing human negotiators with help from outside. The Man-Machine Interaction Group at Delft University of Technology[80] is currently developing a 'Pocket Negotiator'; a device that helps individuals in negotiations by *"increasing the user's capacity for exploration of the negotiation space, reducing the cognitive task load, preventing mental errors, and improving win-win outcomes"*.[81] This pocket negotiator is envisioned to be a smartphone-like device that can be employed in negotiation settings. The Pocket Negotiator is a form of human enhancement, because it may expand and strengthen individuals' capacity to negotiate and to reach better outcomes. Its use, however, may also be contested because it might unduly shift the information position of the parties involved; its use could be perceived as cheating. The Pocket Negotiator gives its user an advantage, both in terms of available information as well as in terms of cognitive capabilities, over the other party.

Neural sensory prosthetics can provide capabilities similar to the Pocket Negotiator, but then (potentially) invisible and embedded in the human body, thereby hiding the information asymmetry from the other party. A bionic attendant of a cocktail party, could for instance obtain useful information

80 http://mmi.tudelft.nl/negotiation/index.php/Negotiation

81 The 'Pocket Negotiator' project proposal, see http://mmi.tudelft.nl/negotiation/images/2/25/Pocket_negotiator.pdf

about others present. Her camera could recognise people's faces, pull up their Facebook profiles and project, unnoticeably to these others, their profile information on her visual implant. This would make her appear very socially attentive and cognizant of the people present. Another use of such implants would be augmented-reality applications similar to those already existing for smartphones. An example is a system that overlays information about houses and their inhabitants over the image shot by the smartphone's camera of buildings in a street. Such information, projected on one's implant, could be useful when talking to a realtor selling the property during a visit to this property. Of course much of this information could have been digested by the buyer prior to visiting the house, but if unexpected things happen – for example, when the realtor suggests visiting another house nearby – the instant provision of information on one's built-in head-up display provides the buyer with an advantage, especially because it is hidden from the other party. Another way of looking at this tilting the information imbalance, is that the buyer restores (or creates) a level playing field. Either way, the method by which the bionic individual improves their information position or power may be hidden from others.

Especially when such applications function without any apparent input from the bionic individual, these applications change the playing field in negotiations and discussions. The information brought to bear by bionically enhanced individuals easily surpasses what other stakeholders in a particular situation may expect of them. In a yard-house sale, sellers may reckon that occasionally a rare expert may turn up who can recognise a piece of value, but they would not expect visitors to come equipped with, for instance, a camera and automated-recognition software that overlays their vision with auction-result web pages. Such bionic enhancement is not necessarily wrong or a form of cheating, but it at least has the capacity to substantially alter social interaction patterns.

Also cochlear implants may provide unexpected advantages in everyday situations. In an international context we may assume that some people speak multiple languages. This occasionally provides for awkward situations when such people switch to their native tongue for a private exchange, only to discover that unexpectedly someone else also understands their language. When equipped with a cochlear implant and a Bable-fish like application, someone could easily turn into a polyglot. Youtube, for instance, already provides simultaneous transcribing audio into text (beta); from there, rough translations in other languages could be provided. Even when the translation is not perfect it may enhance one's foreign language capabilities significantly. When others present in a conversation are not aware of this kind of enhancement on the part of one of the participants, this provides the bionic person with an advantage. Again, this need not be ethically or legally wrong, but it does alter the (expectation of) information balance in social interaction.

Information asymmetry always plays a role in negotiations and other communication environments. Buyers are generally less informed than sellers, which causes market imperfections (Akerlof, 1970). Also, people differ in their capacities to process information and to make choices. Classical economics is based on the assumption of perfect information (and perfectly rational actors). Akerlof (1970) and others (e.g. Stigler, 1961) have shown the effects of imperfect information on the side of consumers; their work stresses that many free-market institutions can be seen as ways of

solving or reducing 'lemon problems'[82] and compensating negative effects of information asymmetries. For instance, insider trading is prohibited in most jurisdictions, mortgage advisors have to disclose their ties with banks and insurance companies, and financial institutions have to provide financial information leaflets outlining the hidden costs of their services. In many of these cases the potentially distorted expectations of one or more involved parties is corrected by additional information to create a more level playing field. Withholding information or bringing information to bear without informing the others involved may be considered foul play if this information is essential for the decisions at hand.

In the same way, using neural prosthetics linked up to external information sources might in certain situations be considered cheating or unethical behaviour. Consider a human-resource manager equipped with the capability of invisibly pulling up information on an applicant, for instance by having her visual support system use facial recognition to retrieve appropriate Facebook information on the applicant. Not informing the individuals under scrutiny of this capacity might be considered unjust, just like people in many countries have to be informed that they are subject to camera surveillance.

It is important to acknowledge that expectations play an important role in determining whether stealthily bringing additional information to bear upon a situation is unethical, and that expectations change over time. For instance, while some years ago it was considered unfair by many human-resource departments to inspect applicants' Facebook profiles, this practice is now less controversial now that every such department is thought to be doing so.

Ethical issues relating to human enhancement

Apart from the ethical implications raised by the problem of information asymmetries of implants, we also face normative issues due to the human enhancement aspect of such applications. As we explained in the introduction, sensory neural prosthetics not only restore or establish hearing or sight of impaired people, but they can also enhance functionality beyond normal sensory perception. In particular, they can be fed by wireless signals that are not in the (human) auditory or visual electro-magnetic spectrums, and thus pick up more information than is possible for people without such sensory neural prosthetics, without the information input being (necessarily) recognisable. It is clear, then, that bionic sensory implants are a form of human enhancement. This in itself, however, does not raise ethical issues; as Savulescu and Bostrom (2009) note, this is only the case if a morally relevant

82 I.e., the problem that buyers with imperfect information take a risk in buying a product that might turn out a 'lemon', a faulty product.

distinction exists between the enhanced and unenhanced functionality. Besides the issue of information asymmetry, which has been dealt with in the previous section, several other aspects seem to merit discussion as they potentially track morally relevant distinctions. We will follow here the general types of argument offered in the literature on human enhancement, applying these to the concrete application of neural sensory prosthetics.

Before we do so, we can set aside some *topoi* in the human enhancement debate that are not particularly relevant to our topic. These are the challenges, usually raised against particular forms of enhancement or against enhancement in general, broadly based on particular normative outlooks: it is against human nature, it is playing God, or it risks changing the human species beyond its intrinsic nature. Or to phrase it less metaphysically and more eloquently, there is a concern with enhancement *"not as individual vice but as habit of mind and way of being"* that reflects an attempt to change *"our nature to fit the world, rather than the other way around"*, and hence, the enhancement mindset of attempting to gain control over ourselves might lead to a loss of *"openness to the unbidden"* (Sandel, 2007, pp. 96-97). Such objections may have value in certain (particularly dignitarian) normative outlooks, although perhaps less obviously so in our outlook of a liberal constitutional democracy, but in any case there is nothing specific in these objections for bionic implants in comparison with the wide range of enhancement technologies. We therefore leave these objections to the debate in the general enhancement literature (see for instance Harris, 2007; Sandel, 2007; Savulescu & Bostrom, 2009). We will also leave aside health and safety issues (except where they touch upon consent issues, see *infra*), and assume that bionic sensory implants do not raise particular health or safety risks. After all, our paper discusses the use of cochlear and retinal implants for information retrieval, and the health and safety risks of this application do not *prima facie* differ from the risks of using cochlear and retinal implants for their primary, medical function of restoring perception. For the sake of argument, we will assume that these medical devices as such are sufficiently safe.

Now we can discuss the remaining ethical enhancement issues as applicable to bionic sensory implants. We distinguish between three major types of argument: the therapy/enhancement distinction, arguments on the level of the individual, and arguments on the societal level.

Therapy versus enhancement

The first issue is discussed and contested in all enhancement debates: is a particular application acceptable for medical purposes (therapy, restoring functionality) but not for other purposes (enhancing functionality beyond 'normal')? Obviously, it is not easy to define what is a 'medical condition' and what is 'normal' functioning, and the grey zone between therapy and enhancement shifts in time and place. We take a pragmatic approach to this conceptual problem, noting that cochlear and retinal implants, as they are developed and used today, generally have a therapy function, aiming to restore or establish sensory perception that is absent or impaired, while the use of these implants for information retrieval, as hypothetically discussed in this paper, generally does not have a medical but rather an enhancement function. (There is an issue whether deafness

should be seen as an impairment or a human characteristic (see e.g., the 'deaf embryo selection' debate in Wilkinson, 2010), but pragmatically speaking, most people with a cochlear implant to restore hearing would consider that therapy rather than enhancement.)

What should concern us here is not the conceptual question, but the material issue whether there is a morally relevant distinction between using sensory implants for therapy or for enhancement. Enhancement advocates would think this is not the case:

> I wonder how many of those who have ever used binoculars thought they were crossing a moral divide when they did so? How many people thought (or now think) that there is a moral difference between wearing reading glasses and looking through opera glasses? That one is permissible and the other wicked? (Harris, 2007, p. 20)

Although Harris has a point here, his rhetorical gusto obfuscates that there may be a moral difference depending on the use of the glasses: if the opera glasses are used not to watch Bryn Terfel on stage but to spy from a distance on Caroline von Hannover sun-bathing on her private yacht, some moral border might well be crossed. Of course, the fact that a technology developed for beneficent purposes might be abused by some for malevolent purposes does not imply that the technology should be prohibited outright (Brownsword, 2009); it could, however, imply that the development or use of the technology should be regulated to control its potential negative uses. For bionic sensory implants, it is therefore relevant to ask whether they are used therapeutically or for enhancing information retrieval and, in the latter case, whether this is morally acceptable in its specific context. Particularly relevant may be the factor of the implant and its use being unnoticed, or unnoticeable, which makes a key difference between therapy and enhancement here. People interact on the basis of other people having normal sensory perception, and they should take into account (e.g., when discussing sensitive issues a little beyond normal hearing distance) that some people have particularly strong hearing, or exceptional visual memory; they will not, however, expect that people they are interacting with have an invisible source of information that directly feeds into their head. Depending on the context, as we have seen in the previous section, this may make a relevant difference to the situation, which will be morally less acceptable if the use of the information stream amounts to cheating in the particular context. There is therefore some reason to believe that, if implants are used beyond therapy, the uses of the implants may need to be regulated depending on their potential for abuse in certain contexts.

Effects on the individual

A second issue, or rather complex of issues, is the potential effect of the enhancement on individuals. A central concern in the enhancement debate is authenticity (Parens, 2009), and attitudes towards enhancement seem to correlate quite strongly with how people perceive the enhancement to affect the authenticity of individual human beings. Enhancement sceptics will argue that artificially enhancing people diminishes their authenticity as human beings; they are no longer true to their ‘real’

self. In a similar vein, Sandel (2007) is worried that our appreciation will vanish of the giftedness of humans: the more human functionality is engineered by enhancement technologies, the less human functioning will be seen as a result of gifts people just happen to have, and this tarnishes the humility that is a key feature of our moral landscape.

Enhancement advocates, however, could retort that enhancement can also improve authenticity, e.g. by widening individuals' range of choices to 'be themselves' or enhancing their intellectual capacity for being authentic.

Whichever stance one takes towards authenticity and enhancement, we see no immediate concern in applying bionic sensory implants for information retrieval. The implants themselves can hardly be seen as diminishing people's authenticity as human beings – we do not regard a pacemaker or a hearing aid as making someone less authentic in some sense, and there is no fundamental difference between these and a sensory implant in this respect. Nor would the use of such implants for external information input seem to make someone inauthentic, i.e., not her 'true' self, at least not as long as the information input is functionally equivalent to standard forms of information retrieval through normal sensory perception.

A shift might occur, however, once this channel of information input would become so important that it starts replacing other forms of information input: there might be an eerie quality to someone who depends largely on unnoticeable forms of perception through implants. This is not a short-term concern – we will be entering the cyborg age once this appears on the horizon. The idea of people, and in particular their brains, being connected wirelessly to external information sources, including the Internet, does not only challenge our notions of authenticity but also of autonomy and identity. If we imagine a world of cyborgs wirelessly connected to each other via the Internet (Warwick, 2002), it is clear we may need to rethink our concepts of what constitutes the autonomy of an individual, as the boundaries of individual body and mind seem to blur in such a world. Also, individuals' sense of self will be affected if their brains are connected wirelessly to external information sources into which they can continuously tap in real time; as people perceive a tool in their hands to be an extended part of their body, so they could perceive an external information source immediately connected to their brain as an extended part of their mind. Possibly, this could lead to changes in the brain's functioning; for example, the brain could adapt to the implants by storing less information in long-term memory while creating more capacity for quickly processing or connecting information.

This sounds like – and is – science fiction today, but a Warwickian world of cyborgs is a possible (if distant) future, and hence we should not discard the consequences of such a development off-hand. It is here that bionic sensory implants do matter, as they can be seen as an initial step on a possible path towards brain/computer-network-interfaced cyborgs. This implies that the consequences of external information sources feeding into the brain for autonomy and identity will have to be taken into account. The short-term implications do not seem particularly significant: as long as the information retrieval remains relatively low-level, individuals' decision-making capacity and their sense of self will not be significantly affected. The mid-term and possible long-term consequences to

autonomy and identity may be larger, however, and we should put these on the agenda for further analysis and discussion.

More immediately relevant, and somewhat more concrete than authenticity and autonomy in general, is the related issue of the appreciation of results of the enhancement: if someone achieves something, for example answering a difficult question, by means of the enhancement technology, does this affect the value of the achievement? A difference seems to exist in our valuation of human functioning: we tend to appreciate an achievement less when it is *"brought about only by means 'separate and external' to the person using them, thus alienating the person from what he or she achieves"* (Merkel, et al., 2007, pp. 341-342). But whether this matters, depends on the context. In some cases, invisible information retrieval may be seen as cheating, particularly in the context of games or other competitive events. Of course, this depends on the rules of the game, as some games or sports allow deceiving the opponent or hiding communications among the team. A useful boundary mark in this respect is whether the deceptive use of neural prosthetics makes use of in-game possibilities or whether it violates the rules of the game. This is similar to the distinction made in virtual-world games in relation to, for example, the theft of a dragon sword: if this is done using in-game possibilities, it is part of the game, but if the sword is appropriated by cracking software code or hacking into someone's game account, it is ethically and possibly legally unacceptable (Kimppa & Bisset, 2005; Lastowka & Hunter, 2004).

Outside of competitive contexts, it also depends on the situation whether invisibly using neural prosthetics would be experienced as somehow 'cheating'. If a Briton's cochlear implant would translate speech from Japanese into English, so that she can understand what is being said, we will value her capacity to understand Japanese less than when she had studied the language for years, in terms of appreciation for her language achievement; but we may also value her ability to interact with Japanese which she otherwise never would have had. If someone answers a difficult question after the answer has been fed into his cochlear implant, we will not appreciate this in the context of a quiz or school exam, but we will appreciate it if we simply need the answer ('what is the antidote for a bite by a red scorpion?'). In other words, it depends on whether the focus of the appreciation is on the process or on the result. In most areas of intellectual performance, what counts will be the results rather than the process of achieving them; only in some areas, notably sports – e.g., chess – and art as well as knowledge testing situations, will the 'authorship' (an achievement of authentic human effort) impact our appreciation of the result (Merkel, et al., 2007). If we conduct a thought experiment that cochlear implants have an embedded functionality to seamlessly translate foreign language(s), in other words act like a Babel-fish, would this affect our appreciation for people's language ability? It would, in the sense that we would no longer admire people for speaking a foreign language. At the same time, however, for most practical purposes it would not matter at all, and if people want to train their brain or distinguish themselves by showing their intellectual capacity, instead of learning Japanese they could try to master quantum mechanics or some other, non-programmable, feat. Hence, except for some contexts in which the process of getting at information matters, such as

sports and knowledge tests, there is no moral objection to information retrieval through implants from the perspective of performance valuation.

A final factor on the individual level is that enhancement can lead to a higher level of individual responsibility: *"As humility gives way, responsibility expands to daunting proportions. We attribute less to chance and more to choice"* (Sandel, 2007, p. 87). Whether this is the case, and whether it matters, is again an issue of context. We can imagine some situations in which more is expected from someone with a bionic sensory implant, for example when she is in a position to provide crucial information in real time (such as the antidote to a red scorpion bite), but in most cases this will not differ fundamentally from someone in a position with any kind of external information source, such as a smartphone. It is only on a more general level that increased responsibility of enhanced people seems relevant; enhancement through implants could reinforce a tendency to point to people's own responsibility for their destiny: why should I tell you something you want to know, when you can look it up for yourself? This is an argument of potentially diminishing solidarity (Sandel, 2007), which no longer resides in the effect of enhancement on the individual, but in the social responses to enhancement.

Effects on society

Here we arrive at the third issue: what are the implications of enhancement through neural sensory prosthetics for society at large? The main issue here is distributive justice. In the enhancement debate, this is typically associated with the question whether the enhancement at issue is an intrinsic or a positional good: does it confer an intrinsic benefit (e.g. a longer or healthier life) or a benefit only in comparison to non-enhanced people (e.g. enhanced height)? Again the distinction is not sharp: certain characteristics, such as intelligence, are positional in some contexts (e.g. in job applications) but intrinsic in other contexts (e.g. being able to enjoy reading Kant), and even typically positional goods may confer an intrinsic benefit in some context (e.g. tallness enabling someone stranded on a desert island to pick fruit from high trees). Enhancement of positional goods raises questions of distributive justice: who has access to them, and who are likely to benefit most? But it should be noted that also enhancement of intrinsic goods triggers such questions: although there may be an intrinsic benefit to such enhancement, if access to the enhancement is unequal, socio-economic inequalities may well be aggravated (Overall, 2009).

How would the enhancement application of bionic sensory implants relate to this issue of *"distributive justice, disadvantaging effects, and the potential for creating an unenhanced underclass"* (Garland, 2004, p. 26)? That would depend first of all on whether people would start having such implants without medical indication, i.e. when they have normal hearing or sight but want enhanced sensory input. This is unlikely to happen in the immediate and perhaps even mid-term future – aside from health and safety issues, the implants are serious interventions in the functioning of the brain, which will need time and effort to adjust to the new form of sensory input. If less-invasive alternatives are available that are roughly functionally equivalent, e.g. miniature hearing aids or glasses with

augmented-reality functions, it is unlikely that people (apart from the likes of would-be cyborgs Kevin Warwick and Steve Mann) will take an implant without medical indication. (It is also questionable whether neurosurgeons will serve healthy people, but that is another issue.) As long as this is the case, there is no special concern with the use of bionic sensory implants; it boils down to the distributive justice of such implants in general. In countries where access to these – expensive – implants is very uneven through inequalities in medical insurance, the disadvantaging effect of implants will be aggravated and there is cause for concern; if on the other hand all impaired people have sufficient access to implants, there is no distributive justice problem.

Suppose, however, that in the more distant future neural sensory prosthetics have become more mainstream and also start to be attractive for unimpaired people. Then the equality issue depends on how costly the implant will be, and which groups are most likely to start using them, and for what purposes. Whether wireless information retrieval through implants serves as a positional or an intrinsic good depends on the context of use; similarly to our discussion above of value appreciation, in competitive contexts (sports, knowledge tests) it would be positional, while in many non-competitive contexts (finding a scorpion-bite antidote) it would largely seem intrinsic. That suggests that there is no reason to generically restrict access to (non-medical) implants per se, but rather to regulate the context-specific use(s) of such implants. There is one situation, however, in which the enhancement implant as such may have to be regulated, namely when only relatively few people from privileged groups reasonably have access to them. In that case, the implant technology as such could aggravate existing social, economic, or cultural inequalities, and if it would be unfeasible to redress the imbalance by subsidising or other facilitating measures for underprivileged groups – which might be too costly for public policy – and it might be more appropriate to restrict access to the implants only for people with medical needs.

Another scenario in the same more distant future is that not a few but very many people start using the implants, to benefit from wireless connections of the brain to external data sources. Then the nature of human interactions could change significantly, perhaps radically, particularly if the external sources would be interconnected in a Warwickian cyborg network. One could argue, as Leon Kass does for life-extension enhancement, that *"the cumulative results of aggregated decisions (...) could be highly disruptive and undesirable, even to the point that many individuals would be worse off through most of their lives"* (Kass in Brownsword, 2009, p. 135). But as Brownsword (2009) rightly points out, this is hardly compelling in the absence of any realistic basis to estimate and balance the potential beneficent and disruptive consequences. Technologies continuously change social practices, not seldom radically – as with the printing press, the telegraph, and the mobile phone – but change in itself does not provide a basis for ethical concern. If human communication and interaction patterns change through bionic sensory implants when many people desire to use them for enhanced information retrieval, there will be a need for close monitoring of potential negative consequences, but not for outright or generic precaution. *"The right thing to do is to make as many better as we can, not to make no-one better"* (Brownsword, 2009, p. 136).

This discussion of arguments at the societal level shows that there may be reasons to regulate the enhancement use of bionic sensory implants, but these reasons point to fine-tuned, responsive rather than generic, command-and-control forms of regulation. A precautionary approach aiming to curb the access to or use of bionic sensory implants for enhancement purposes would amount to overkill and a disproportionate limitation of the benefits these implants could have for individuals. Apart from the principled reasoning, also practical arguments oppose a generically negative regulatory tilt: as with many reproductive or other types of enhancement technologies, absent a global consensus that is very unlikely to emerge, people could easily travel abroad to acquire an implant. These would presumably belong to the privileged classes that are able to afford such 'implant shopping', which is another reason why the distributive justice argument suggests a responsive regulatory approach focusing on compensatory measures that ensure that the enhancement is sufficiently accessible to all, that its use does not violate rights of others, and that the enhancement does not harm the infrastructural conditions of a moral community (Brownsword, 2009; Overall, 2009).

Legal issues

Many legal issues are relevant for bionic sensory implants, but few of these seem to be very specific for this particular application. Most issues are equally relevant for implants in general. For example, the right to bodily integrity (see e.g. art. 3 Charter of Fundamental Rights of the EU) is a key issue for human implants, and informed consent plays a crucial role in exercising this right (Beyleveld & Brownsword, 2007). Patients must be given sufficient information, in a language they understand, about the health and safety risks associated with implants. The risks of bionic sensory implants are not yet well-known, particularly for the embryonic visual and brain/computer-interface implants (Merkel et al., 2007). Practitioners and patients will therefore be very cautious to use implants unless there is a serious medical reason; the non-medical or recreational use of bionic implants resides in a more distant future. Once that future arrives, other legal questions will have to be addressed that are familiar from the enhancement literature, such as whether medical practitioners may implant bionic sensors without a medical indication, and whether employers (e.g. in the military) may force or nudge their employees to take an implant for non-medical reasons. These questions can be left aside for the purposes of our paper, since they are not specific to the use of bionic sensory implants for information retrieval purposes. Perhaps two aspects may be specifically relevant, however. The first is that the use of bionic sensory implants for information retrieval, if used on a structural and longer-term basis, may affect the structure and functioning of the brain perhaps in other ways than cochlear or retinal implants for 'normal' hearing or seeing do. The brain is known to be plastic and could therefore adapt to structural changes in information input. This could have unknown side-effects that will need to be studied, in order to provide would-be implantees with sufficient information about risks and effects to allow them to form informed consent. Another issue is that with information being input into bionic sensors, the likelihood is going to increase that also unwanted or unexpected information is going to be input. In particular, the prospect of computer viruses infecting bionic implants should be studied

carefully. This is less science fiction than it may sound: experiments have already shown the possibility of infecting a human Radio Frequency Identification implant with a computer virus (Gasson, 2010). Hence, wirelessly connected information-processing implants must be developed and researched with great caution and particular attention to abuse and malware.

Another cluster of legal issues relates not to implants as such, but to information processing and the associated advantage that this type of human enhancement may carry. Here, typical legal issues in information law will apply, such as consumer protection and data protection, as well as sectoral legislation applying to information in particular contexts, such as employment or education. These are also not very specific to bionic sensory implants; for example, data-protection norms will apply equally to processing personal data (e.g. googled through a wireless Internet connection) on a pair of augmented-reality glasses as to processing personal data in a bionic implant. One area where this type of implant may have some specific thrust, however, is sports. Sports regulations apply different standards to using headphones for athletes to be connected to their coaches; while this is regular practice in for example cycling, it is prohibited in other sports, such as soccer. Naturally, it is also not allowed in brain sports that depend on the athlete's information-processing capacity, such as chess or go. Should bionic sensory implants become in much wider use in the future, each sports area will have to assess whether they have to adapt their regulations to this development. Somewhat more directly relevant – although the technology is still embryonic – is the case when the prospect of bionic sensors implanted for medical purposes could provide some sort of compensatory advantage. Suppose that at some point retinal implants will allow blind or poor-sighted patient to retrieve a substantial part of their eyesight, while at the same time enhancing some sight-related functionality such as eye-hand co-ordination. In that case, some exceptionally gifted people with implants could aim to participate in regular rather than Paralympic competition, for example in archery or biathlon. The case then would be similar to Oscar Pistorius, the 'blade runner' with two carbon-fibre transtibial (i.e. below-the-knee) prostheses who was ruled ineligible by the International Association of Athletics Federations to participate in regular competition, as his prostheses were thought to confer a considerable advantage over athletes without such prostheses. However, the Court of Arbitration for Sport (CAS) overruled the IAAF decision, arguing that there was insufficient evidence of an overall net advantage in Pistorius's case.[83] For neural prosthetics used by impaired athletes, similar cases may arise in the future that will call for a (case-by-case) assessment whether they bring an overall net

[83] Court of Arbitration for Sport (CAS) 16 May 2008, case 1480 (*Pistorius v. IAAF*), see http://jurisprudence.tas-cas.org/sites/CaseLaw/Shared%20Documents/1480.pdf

advantage compared to people without such implants; as the CAS (2008, para. 56) observed in the Pistorius case, this is *"just one of the challenges of 21st Century life"*.

Conclusion

In this paper we have discussed a particular kind of human enhancement, a class of prosthetics that enhance human (auditory/visual) sensory capabilities and which, because auditory and visual tracts convey information, also indirectly enhance human cognitive capabilities. These implants derive their value not from enhancing the cognitive apparatus ('computing power'), but from providing more and better input for processing ('content'). External intelligence, for example in the form of smart applications that run on external devices, can tap relevant information directly into the nervous system and hence augment other signals entering this system. While similar effects in terms of bringing relevant information to bear can be achieved with external mounted displays and ear pieces, the implants are potentially more disruptive in human relations because their presence and functioning may be hidden from other stakeholders present. The implants could, if visible from the outside, also be mistaken for therapeutic devices, such as hearing aids, and thus solicit a sense of sympathy rather than vigilance. When undetected by other stakeholders, these sensory neural prosthetics might uneven the informational playing field to an extent that it becomes foul play. Neural prosthetics are particularly challenging in situations that are crucially dependent on information positions, such as negotiations, knowledge tests, or sports. This could warrant, for example, notification obligations on the part of the bionic human or other types of context-specific regulations aimed at compensating for the informational advantage of bionic implants.

As we have shown, the effects of bionic sensory implants go beyond the informational plain. In the longer term, if the implants become more prevalent and if people start to use them as a consistent source of information input, the processes responsible for the information feed to the implants will affect how bionic people think, feel and behave. For example, if the implants constantly provide information on top of, or replace, information perceived by the bionic individual, this may literally lead to tunnel vision, and selective information input or processing could reinforce cognitive biases in bionic humans as well. At the same time, the broader scope of information sources may also widen people's point of view, and it could counter-balance cognitive biases by alerting people to information that their own sensory perception fails to notice. It all depends on how the information feed and the brain are going to interact.

On a more fundamental level, the long-term scenario also has the potential to significantly affect human autonomy, identity and authenticity. Implants with a consistent external information feed could well be perceived as an extended part of the human body or human being, and because of the wireless connection of neural prosthetics, this challenges our notion of the boundaries of a human being more than is the case with physical technological extensions of the human body. Bionic sensory implants can be seen as a first step towards a future scenario of Internet-connected, and perhaps

mutually interconnected, cyborgs. These potential long-term effects call for reflection on how these implants can and should be developed in the short and middle term.

There is no need, however, to be overly precautionary in applying bionic implants, which already have significant therapeutic benefits and which may have equally interesting non-therapeutic benefits in the future. The fact that they also have potential for malevolent use and possible unknown side-effects should not lead us to an overall restrictive regulatory tilt. The vision of bionic human beings will not appeal to all and perhaps be scary for a majority of people today, but it would be moral arrogance for us to draw a line of allowing sensory implants only for purely therapeutic reasons. Indeed, it would be moral arrogance to presume we know what is best for future generations based on our current outlooks (Hanson, 2009). As Brownsword (2009) reminds us:

> We should not forget (...) that ethical objections to enhancements are not the whole story; even if an enhancement is morally permissible, it does not follow that we should welcome it; but neither does it follow – and this, I think, is the fundamental point – that we have a right to impede the morally permissibly simply because we do not welcome it (Brownsord, 2009, p. 152).

Rather than command-and-control regulation with a negative tilt, we should therefore closely monitor the development of neural prosthetics and discuss their ethical and legal implications on a timely basis. To prevent cheating with bionic implants while fostering their fascinating information potential, we recommend a responsive regulatory approach. This should focus on context-specific compensatory measures that ensure that neural prosthetics are sufficiently accessible to all, that their use does not violate rights of others in specific situations, and that they are in line with the infrastructural conditions of a community of human beings centering on social interaction.

References

Adams, D. (1980). *The hitchhiker's guide to the galaxy* (1st American ed.). New York: Harmony Books.

Akerlof, G. A. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, *84*, 353-374.

Beyleveld, D., & Brownsword, R. (2007). *Consent in the law*. Oxford: Hart.

Bostrom, N., & Savulescu, J. (2009). Human enhancement ethics: The state of the debate. In J. Savulescu & N. Bostrom (Eds.), *Human enhancement* (1-22). Oxford: Oxford UP.

Brindley, G. S., & Lewin, W. S. (1968). The sensations produced by electrical stimulation of the visual cortex. *The Journal of Physiology, 196(2),* 479–493.

Brownsword, R. (2009). Regulating human enhancement: Things can only get better? *Law, Innovation & Technology, 1(1),* 125-152.

Court of Arbitration for Sport (2008), case 1480 (*Pistorius v. IAAF*). Retrieved from http://jurisprudence.tas-cas.org/sites/CaseLaw/Shared%20Documents/1480.pdf

Foucault, M. (1978). *Surveiller et punir. Naissance de la prison [Discipline and punish. The birth of the prison].* Paris: Gallimard.

Garland, B. (2004). Neuroscience and the law. A Report. In B. Garland (Ed.), *Neuroscience and the law. Brain, mind, and the scales of justice* (1-47). New York, Washington, D.C.: Dana Press.

Gasson, M. N. (2010, 7-9 June 2010). *Human enhancement: Could you become infected with a computer virus?* Paper presented at the 2010 IEEE International Symposium on Technology and Society, Wollongong.

Hanson, R. (2009). Enhancing our truth orientation. In J. Savulescu & N. Bostrom (Eds.), *Human Enhancement* (357-372). Oxford: Oxford UP.

Harris, J. (2007). *Enhancing evolution: the ethical case for making better people*. Princeton, NJ: Princeton University Press.

Kimppa, K.K, & Bissett, A.K. (2005). The ethical significance of cheating in online computer games. *International Review of Information Ethics, 4, 31-38.*

Lastowka, G., & Hunter, D. (2004). Virtual crime. *New York Law School Law Review*, *49*, 293-316.

Merkel, R., Boer, G., Fegert, J., Galert, T., Hartmann, D., Nuttin, B., Rosahl, S., & Wuetscher, F. (2007). *Intervening in the brain: changing psyche and society*. Berlin, New York: Springer.

Moore, C. W. (2003). *The mediation process: Practical strategies for resolving conflict* (Vol. 3). San Francisco: Jossey-Bass.

Neale, M. A., & Bazerman, M. H. (1991). *Cognition and rationality in negotiation*. New York: The Free Press.

Overall, C. (2009). Life enhancement technologies: The significance of social category membership. In J. Savulescu & N. Bostrom (Eds.), *Human enhancement* (327-340). Oxford: Oxford University Press.

Parens, E. (2009). Toward a more fruitful debate about enhancement. In J. Savulescu & N. Bostrom (Eds.), *Human enhancement* (181-197). Oxford: Oxford University Press.

Sandel, M. J. (2007). *The case against perfection: ethics in the age of genetic engineering*. Cambridge, Mass.: Belknap Press of Harvard University Press.

Savulescu, J., & Bostrom, N. (2009). *Human enhancement*. Oxford; New York: Oxford University Press.

Simon, H.A. (1947). *Administrative behavior: A study of decision-making processes in administrative organizations.* New York: The Free Press.

Stanley, G. B., Li, F. F., & Dan, Y. (1999). Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *The Journal of Neuroscience, 19(18),* 8036-8042.

Stigler, G. J. (1961). The economics of information. *Journal of Political Economy, 69 (3)*, 213-225.

Tancredi, L. R. (2004). Neuroscience developments and the law. In B. Garland (Ed.), *Neuroscience and the law* (71-113). New York, Washington, D.C.: Dana Press.

Thompson, L.L. (2005). *The mind and heart of the negotiator*. Upper Saddle River, N.J.: Pearson/Prentice Hall.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453-458.

Warwick, K. (2002). *I, cyborg*. London: Century.

Warwick, K., Gasson, M., Hutt, B., Goodhew, I., Kyberd, P., Andrews, B., Teddy, P., & Shad, A. (2003). The application of implant technology for cybernetic systems. *Archives of Neurology, 60(10),* 1369-1373.

Wessberg, J., Stambaugh, C. R., Kralik, J. D., Beck, P. D., Laubach, M., Chapin, J. K., Kim, J.,Biggs, S. J., Srinivasan, M. A., & Nicolelis, M. A. L. (2000). Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature, 408(6810),* 361-365.

Wilkinson, S. (2010). *Choosing tomorrow's children: the ethics of selective reproduction*. Oxford: Oxford University Press.

Chapter 10
Why should I be natural? A fivefold challenge to the supposed duty to 'be natural' as grounds for outlawing human enhancement

Pieter Bonte
Ghent University
Bioethics Institute Ghent
✉ pieter.bonte@ugent.be

Abstract Human enhancement technologies put us at liberty to *materially* remake ourselves from our appearance over our physical capacities right down to our own mental activity: we are becoming 'self-shaping animals' through and through. Proclaiming a moral duty not to transform our human nature – a duty to be and remain an unmodified *homo sapiens* – 'bioconservatives' consider the enhancement enterprise *unnatural* and *dehumanizing,* to be condemned and according to some even outlawed. This paper denies that such a duty exists by adding the following arguments to the often heard *naturalistic fallacy* objection: (1) human nature may come to hinder our pursuit of worthy goals, (2) evolution made us 'crooked timber, out of which no straight thing can be made', if not for the help of enhancement technologies, (3) even if we could consider ourselves 'well-created beings', nothing should hold us back to ennoble ourselves even further, (4) our default biological determinations are in principle even more estranging to us than the insertion of artefacts in ourselves, and finally (5) in principle, there is greater unfairness in the 'natural lottery' than in a well-guided policy of 'enhanced equality'. This battery of arguments deeply undermines the belief that there is a 'duty to be natural'. However, this does not imply that human enhancement *ipso facto* becomes a laudable undertaking, nor that human nature should be ignored altogether when deciding how to improve our lot. It does, however, discredit the notion that there are ethical reasons to categorically conserve our *homo sapiens* nature, come what may.

Keywords human enhancement, human nature, human dignity, nature's normativity, self-determination

Nature's normativity: flogging a dead horse, with a twist

> The words ['nature' and 'natural'] have come to excite…feelings which…have made them one of the most copious sources of false taste, false philosophy, false morality, and even bad law (John Stuart Mill, 1904).

This paper attempts to address the *"deeper infrastructural concerns"* (Brownsword, 2009, p.134) raised by the possibility of fundamentally altering human nature and the ensuing *"feelings of uneasiness, creeping disorientation and even existential panic"* (Gordijn in Brownsword, 2009, p.

134). Slowly or quickly but in any case surely,[84] we are materially re-forming our own bodies from brain to toe, welcoming adequately safe and effective 'human enhancement technologies' (henceforth HETs) in our bodies, right down to our brains. These HETs allow us to manipulate and 'enhance' ever more aspects of our own appearance, physical ability, cognitive performance and emotional condition. Such enhancement practices have come to pervade Western societies, yet they generate significant discontent. Many fail to rejoice in this 'liberation biology', as some label it[85]*:* every household watching *Extreme Make-Over*, every sports fan seeing doped athletes disrupt his beloved game, every troubled soul staring down at his bottle of psychoactive drugs – surely many, if not all of them, have asked themselves the vexing question: *shouldn't we stay natural in all of this?*

Philosophical parsing of this acutely felt need to 'be natural' in the face of HETs has been intensive the past decade or so, and very informative.[86] More often than not, such parsing ends up thoroughly refuting the idea that our biological nature can be a source of normativity (Buchanan, 2009a; Caplan, 2006; Dennett, 1992; Hughes, 2009; Pinker, 2002). The disillusioned understanding of ourselves as beings who are biologically emergent from pointless, pitiless natural selection, brings these authors to the moral conclusion that, in direct contradiction to the fuzzy intuition to 'stay natural', we ought to feel *morally estranged* from that nature on account of its sheer pointlessness and pitilessness. Modern day evolutionists voice, in the harshest of tones, their *reasoned repugnance* towards the amoralities of *'Nature, red in tooth and claw'*.[87] Perhaps Dawkins (2006), whose adversaries never fail to misrepresent his opinion on this issue, speaks most clearly here when exclaiming in *The Selfish Gene* that we must *"rebel against the tyranny of the selfish replicators"* (Dawkins, 2006, p. 201).

Whether we *de facto* remain deeply influenced by our evolutionary heritage can be a point of intense and important debate[88], but at the heart of that dispute, both 'social constructivists' and 'natural determinists' share the moral conviction that nature must never in itself be taken as a guide

---

[84] Expert opinions differ widely on the speed of current and expected technological advance. Compare, for example, the highly educated guesses of Pinker (2003) and Silver (2006).

[85] For example, libertarian science writer Ronald Bailey (Bailey, 2005).

[86] For instance, two collections of papers on precisely this topic have recently been published in book form: '*Is human nature obsolete?'* (Baillie & Casey, 2005) and '*The normativity of the natural'* (Cherry, 2009).

[87] Which they share with the social constructivists but to whose world view they (purport to) add a firmer, clearer grasp of natural science.

[88] For a thorough round-up of the contemporary state of the 'nature/nurture' debate, see Pinker (2002).

on what is right and wrong.[89] From this shared perspective, then, our dignity, duty and only hope for finding true purpose in life would lie in establishing *"an identity which lies at an ever-increasing distance from our organic nature"* (Tallis, 2007).

Despite this broad consensus on the non-normativity of human nature, more recent reflection on HETs has sparked a resurgence of 'natural law' theories that quite literally advocate a principled duty to cherish and conserve our *homo sapiens* constitution (Annas, Andrews, & Isasi, 2002; Fukuyama, 2002; Kass, 1997, 2003; President's Council on Bioethics, 2003; Sandel, 2007; UNESCO, 2005). At the same time, out of the underlying consensus of social constructivists and natural determinists, new moral and political philosophies are sprouting up, unified by the belief in a principled right (or in stronger versions, even a *duty*[90]) to enhance ourselves. These new philosophies frame human enhancement as a program of systematically shedding all our unwanted natural atavisms through civilisation. For such a project of civilisation, our conventional means of moral education and the creation of wholesome environments can already achieve a lot, but, according to the proponents of enhancement, only the 'civil engineering' of ourselves with HETs will ever take us all the way.

In this paper, I will add five arguments to the basic 'naturalistic fallacy' critique. These arguments profoundly discredit the belief that we 'ought to be natural'.

[89] Upon gaining a rudimentary evolutionary self-understanding and having witnessed the horrific consequences of political systems acting on a supposed 'duty to be natural', post-war social constructivists, for instance, went out of their way to *exclude* mankind from natural determination: some fortunate quirks of evolution had allowed us to become *cultural creatures* all the way down, they contended. The happy upshot was that the beastly ethic of survival of the fittest was only loosely nestled in us, or perhaps even not at all. In any case it could be wholly supplanted by moral laws of our own devise, and precisely such full moral autonomy, completely unfettered by nature, was celebrated as the best thing that ever happened to us (Pinker, 2002). According to the current scientific consensus, although certain parts of our biological constitution are culturally flexible and susceptible of further improvement, others have a tenacious rigidity to them and will continue to manifest themselves despite our best efforts to eradicate them.

[90] For human enhancement as a *moral* duty, see Dworkin (1999), Harris (2007), Bostrom (2008); for a novel view on how human enhancement might even come to be seen as a something that is to be promoted and perhaps mandated through law (a 'state eugenic' prospect which 'liberal eugenicists' such as Agar (2004) try to evade as much as possible), see Buchanan (2008).

Human nature on the stand: a fivefold challenge

> To entrust oneself to Nature, that is to entrust oneself most wisely. Oh! What a soft and sweet resting pillow it is (de Montaigne, 2004).

A host of divergent debates can be started on what is to be understood by 'human nature'. We do not need to go into those matters too deeply here. Instead we will take as a vantage point the carefully crafted definition of human nature provided by Buchanan (2009b) in *'Moral status and human enhancement'*, which I dub 'recalcitrant homo sapiens nature' (henceforth RHSN):

Human nature is a set of characteristics:

1. that most beings that are uncontroversially human have at this point in biological and cultural evolution (and have had throughout what is uncontroversially thought to be human [as opposed to prehuman] history);

2. that are relatively recalcitrant to being expunged or significantly altered by education, training, and indoctrination; and

3. that play a significant role in explanations of widespread human behaviour and in explanations of differences between humans and other animals (Buchanan, 2009b).

This definition has the benefit of zooming in on that proportion of ourselves which, by definition, is not easily malleable via the conventional cultural manipulations of education, training and indoctrination.[91] Whether that proportion is substantial or negligible is an empirical question that has received an empirical answer: humans are certainly no 'blank slates', open to infinite perfectibility via the relatively 'easy way' of cultural manipulation, as was once widely perceived to be the case (Pinker, 2003). A RHSN resides in us, for the civilisation of which conventional cultural manipulation is relatively impotent. To the extent that it is civilisable, such civilisation will only be possible via the unconventional means of HET. So the question asked here is not: Should we prefer conventional cultural means? The question is rather: Should we aim to civilise our RHSN, or are such attempts misguided and should we instead wish to conserve our RHSN? A separate follow-up question is: If civilisation of our RHSN is desirable, should we use the unconventional means of HET?[92]

---

[91] As such it straightforwardly acknowledges being a strongly 'theory laden' definition, designed specifically to sharply distinguish these recalcitrant biological features from the culturally plastic ones.

[92] This question needs to be separated from the previous one, because it is logically possible that our answer to

To answer these questions, I will raise five challenges to the idea that we should conserve our RHSN.

Human nature can hinder us in pursuing worthy goals

Today perhaps more than ever, we have a clear and precise awareness of how our RHSN hinders us from achieving worthy goals, such as cognitive and athletic prowess, two fields of human agency we will highlight here. What is more, RHSN not only hinders us in such a way that it may *slow* us or may *make it hard* to achieve these worthy goals, rather in some cases it literally *blocks* us and *makes it impossible* to achieve those goals. Whenever this is the case, we then have to offset the value we give to the conservation of our limited *homo sapiens* nature against the value of pursuing those goals by enhancing our RHSN.

Cognitive stagnation

Without the prospect of us developing cognitive enhancers, it is arguable that we are reaching certain limits of our intellectual faculties, a thesis explored – not without added journalistic dramatism – by a host of scientists and philosophers in John Horgan's '*The end of science: Facing the limits of science in the twilight of the scientific age'* (1997). For how many minds, if any, can truly master the staggering complexities of quantum physics? And, how many minds, if any, can truly master a unified comprehension of *all* fields of science and all of the humanities combined? No level-headed, modern-day intellectual would even begin to think this possible: such a humanistic *uomo universale* ideal has long been buried. Nevertheless, such a 'consilient' understanding and wisdom is what many people hold to be humanity's deepest and most worthy cause – an ethic famously expounded in E.O. Wilson's contemporary classic '*Consilience – The unity of knowledge'* (1999). From this perspective, only the advent of effective cognitive HETs could begin to remedy the deep cognitive inabilities we have had to diagnose in our species, and this prospect has begun to spark new optimism among certain intellectuals.[93]

---

the first question would be that we should aim to civilize our RHSN, only to find there being something unethical in the means of HET to such a degree that we should, somewhat tragically, refrain from using such means, even as we accept that our RHSN should be amended *in some other way* than through conventional cultural manipulation (for that has proven too weak) or HET (for that has proven unethical).

[93] See for instance Hughes (2004), Harris (2007), Buchanan (2008), Bostrom (2008). Also interesting to note is the poll *Nature* took amongst its readers, revealing that some researchers are already 'brain doping' themselves with the crude and limited cognitive enhancers available today (Maher, 2008). No strong conclusions can be

Athletic stagnation

If the current *categorical* bans on sports doping persist in spite of the strong critiques of certain ethicists (Caplan, 2009; Kayser, Magnon, & Miah, 2007; Kious, 2008; Tännsjö, 2009), we will have to face up to the fact that we thereby intentionally call a halt to the historical process of *record setting*.[94] In contrast to all previous centuries, we are today effectively reaching the peak performance of our homo sapiens bodies in many respects, such as flexibility, running speed, lifting power, jumping distance, endurance, etc. [95] We have virtually perfected our methods of scouting, training and dieting, and we cannot take our unenhanced bodies much further: biologically, we are at the end of our rope – a conclusion drawn in several recent scientific studies, one of which predicts that the end point of meaningful improvement of the athletic potential of the human body will be reached some fifteen years from now, around 2027 (Berthelot et al., 2008; Lippi, Banfi, Favaloro, Rittweger, & Maffulli, 2008).[96] Therefore, if doping continues to be categorically banned, in due course further records will only spring from the use of new equipment or changed rules, and our sporting culture will effectively lose its 'epic' dimension of making the human body itself go ever 'swifter, higher, faster', as the Olympian

---

drawn from that poll, however, as it was not based on a scientific questionnaire methodology, but other, more stringent inquiries into the popular use of cognitive enhancers have been undertaken too. For a quick review thereof, see Talbot (2009). Besides such principled love of wisdom, there are also strong prudential arguments urging us to enhance ourselves, *conservative* arguments even. As Buchanan (2008) points out, cognitive enhancement may be required for the conservative goal of maintaining sustainable, peaceful societal structures that prove successful in warding off incumbent environmental and geopolitical disaster. Indeed, a conservation ethic soberly aimed at 'just keeping things as they are' may ironically require us to implement a plethora of HETs in our bodies, cognitive and non-cognitive, from improving our capacity to extract nutrients from foodstuffs to increasing our impulse control, altruistic sentiment or cognitive capacities such as better foresight or a better understanding of the implications of probabilities and proportionalities (Buchanan, 2008).

[94] One may consider 'record setting' to have no value in itself, or that, as soon as our natural potential has been stretched as far as possible, other cultural conceptions of sporting should then trump a further pursuit of absolute peak performances.

[95] The critically acclaimed documentary '*Bigger, faster, stronger – The side effects of being American'* (Bell, 2008) provides many unsettling testimonials of the transgressions of this natural threshold.

[96] "The proposed…model…suggests major global fading of world record progression…In all measurable Olympic contests from five different disciplines, involving either aerobic (10000 m skating) or anaerobic (weight lifting) metabolic pathways, leg muscles mainly (cycling) or all muscles (decathlon), lasting seconds (shots) or hours (50 km walk), either in men or women, small (Fly weight) or tall athletes (100 m free style), individual or collective events (relays), all progression curves follow the same pattern" (Berthelot et al., 2008, p. 4).

motto would have it. No longer will our sports arenas be fora for truly unique, historically unseen peak performances with universal resonance. And as this absolute, epic dimension of sport withers, sport is reduced to its relative, 'dramatic' dimension of winning and losing against directly present competitors. 'Citius, altius, fortius', taken literally, seems to require *allowing doping*, certainly from 2027 onwards. (As to the contention that doped performance is not the same as 'natural performance' because a doped athlete did not perform his athletic feat *himself*, I will later in this paper show that this argument is only superficially convincing.) *Banning doping* seems to require the Olympic motto to be erased and replaced, or at least be given a new, non-literal interpretation.[97]

Confronted with such cognitive and athletic limitations, apparently insurmountable by conventional means, one may think it wise to counsel resignation, acceptance and contentment with what is within our naturally circumscribed reach. However, resignation and contentment can hardly be proposed as *basic* virtues, for if stressed too strongly, they can degrade into *fatalism* or sheer *laziness*

[97] Focusing on the peculiar competitive nature of elite sports can give rise to three further criticisms of the use of HETs. Firstly, because sports is often a *competitive* endeavor, it can be said that the goal – the mere beating of an opponent – is hardly something intrinsically worthy, and so using HETs to further pursue it is no worthy undertaking neither. But if *that particular* goal of 'beating others' is not worthy, that still leaves open the possibility of there being *other* athletic goals that *are* worthy, such as self-expression and improvement, creative and progressive experimentation, expanding the existential boundaries of mankind, etc. – and HETs can be used to pursue these goals too. Secondly, its competitive nature can be grounds for denying competitors the right to dope on account of it being *unfair.* I hope to counter, perhaps even *invert* this argument from unfairness in this paper. Thirdly, elite sports delivers *'positional* goods': if a few athletes start using doping, this may raise their competitive advantage and as a result they obtain higher rankings. But if *all* athletes would use doping, then all advantages would be lost, and everyone would be back in the same position, only now they all dope – with all possible risks that entails – only not to *lose* that position. Therefore, it seems better for all if it is agreed and ensured that no one starts to use doping. This certainly is a sobering thought, but two counterarguments have to be taken into account. Firstly, different athletes will react differently to each doping means: some physiologies will react better to it than others. Therefore, it seems unwarranted to *actively forbid* the use of such means for personal advancement to those that would benefit *more* from it, for that would be forcing a static population perspective on every individual's life perspective. Secondly, even if the individual goods obtained by athletes remain profoundly positional – there can always be only one number one – the aggregate effort of striving to be that number one does raise the entire athletic enterprise to new and higher levels of performance. However, a more fundamental problem still slumbers, as it is disputed whether a *performance* based on the use of HETs can be considered to be a true *accomplishment,* properly attributable to the athlete. *This* argument is dealt with later in this paper. I want to thank an anonymous reviewer for addressing this added complexity of competition in the domain of athleticism.

– the unjustified wasting of human potential. Resignation and contentment are most reasonably seen as *balancing* or *reactive* virtues, as they function in the famous *Serenity Prayer: "God, grant me the serenity to accept the things I cannot change, courage to change the things I can, and wisdom to know the difference."*[98] Serenity lies in knowing you have done all you ethically can to achieve worthy goals, given your circumstantial limitations.

Today, HETs are mostly still too unsafe, too inefficacious or too costly for it to be wise to advocate their use, even in the pursuit of worthy goals. So in that sense, it may be advisable, anno 2011, to find serenity in resignation. In the same vein, we do well to resist the millennial tendencies of those 'transhumanist' philosophies that envision today's humanity to be *"the future living in the past",* thus demonstrating an escapist urge that can easily radicalise into an attitude of disregard and contempt for our current predicament, as Hughes (2007) – himself a dedicated but mindful transhumanist – acknowledges.

Be that as it may, by the same 'logic of serenity' *the research and development of* efficacious, safe and cheap HETs may well be a moral duty that is already incumbent upon us, although it may have to be assigned a rather low priority when considering the many basic health care policies we still fail to provide to an adequate degree. But no matter how such a complicated calculus of political priorities may turn out: to the extent that we forestall enhancement research without proper overruling priorities, future generations may feel bereft, perhaps dramatically so, and reprehend us for forestalling the benign spread of HETs through society.[99]

---

[98] The origins of this aphoristic prayer are disputed, but most ascribe it to 20th century theologian Reinhold Niebuhr, see Goodstein (2008).

[99] As far as *rights* are concerned, anyone who takes issue with the argument here presented and does not consider RHSN as 'stunting' but finds a way to consider it a 'proper proportioning' of human capabilities, should surely be allowed to abstain from using HETs. But they might do well to prepare themselves for a scenario in which they could become not only economically surpassed by the more bioprogressive cultural groups, but profoundly culturally isolated as well. For instance, McKibben (2004), an outspoken anti-enhancement environmentalist, acknowledges that resisting HETs may involve taking a stand somewhat similar to the one taken by the Amish today in resisting much of 20th century technology. Yet to a 'bioprogressive' ethic, such a submission to nature may ultimately be seen as a self-afflicted amputation, a denial of fundamental human duties to improve our understanding of the larger world as it really is, to accept full responsibility for ourselves and create the deepest, richest life possible. This goes to show the extent to which the enhancement debate may bring about a fundamental 'culture clash' that may prove difficult to pacify.

Nature made us 'crooked timber'

The comments I made in the previous section still allow us to consider our RHSN as being limited, but nonetheless fundamentally *good*, and perhaps we would do best to refrain from trying to fix things that are not really broke to begin with, such as our RHSN – 'Le mieux est l'ennemi du bien'. Evolutionary science, however, has proven such views to be false, because our RHSN really is 'broke to begin with' in certain crucial respects.

Firstly, the processes of natural selection that shape human biology have a significant developmental drawback. Roughly said, evolutionary adaptations can only come about in long processes made up of many incremental steps, and in every incremental step, the entire existing organism and all its interdependencies have to be taken into account. Adaptations are always 'made up as things go along', via the universal dynamic of random genetic mutation followed by natural selection through environmental pressures – a pseudo-engineering process aptly captured in the anthropomorphic metaphor of *'The blind watchmaker'* (Dawkins, 1996). As a result, organisms are replete with 'evolutionary trade-offs', where the best possible optimum often involves significant compromise. The same is true for humans. Purely physical examples are our notoriously bad backs, and our notoriously slim pelvises, making human childbirth such an excruciating ordeal in comparison to much of the animal world. There is no 'deeper meaning' to the extreme labour pains suffered by human mothers, just like there is nothing to be revered in all the other unhappy evolutionary compromises we carry within us. Arguably, such evolutionary compromises can in principle be alleviated by our own engineering – it will frequently be incredibly difficult, but given enough time, it may also often prove possible. In the case of childbirth, we already know the 'enhancements' of the caesarean section[100] and epidural anaesthesia, and perhaps we should long for the possibility of outright extracorporeal pregnancy (ectogenesis) in artificial wombs or 'human hatching eggs', which would make humans join the ranks of the platypus as one of the few egg-laying mammals still alive on earth.

Secondly, every human being born today is genetically quasi-identical to the homo sapiens that spawned some 200.000 years ago. The bodies and basic mindsets we develop today are adaptations to our 'environment of evolutionary adaptedness' or EEA, which is, roughly said, the savannah

[100] The example of the caesarean section only goes to show that innovations that enhance our biological predicament in some respect, may be deeply intertwined with adverse effects that possibly eclipse all enthusiasm over the enhancing effects. In the case of the caeserian section such effects include risks to future fertility and a host of other crucial side effects both in the short and the long term. See for instance http://guidance.nice.org.uk/CG13 (accessed 6th of January 2011).

wilderness and stone age culture we were surrounded with 200.000 years ago. Since then we have experienced massive changes in our cultural environment, but our genome has not changed with it. As a result, our homo sapiens bodies and basic mindsets have become very *maladapted* to some of our cultural surroundings. A classic example of this genome-culture mismatch is our hard-to-control craving for sweet and fatty foods. We crave these nutrients so crazily, because in our EEA they were very scarce yet very nutritious. Therefore, evolutionary processes have made us very motivated to find them and, once found, to gorge them down. Advanced agriculture has since long put us in the odd spot of having access to an overabundance of sugars and fat. As a result, we now have to fight back our prehistoric craving for them, a feat which only few manage without frustration. Now, as it is with the pains of childbirth, so it is here: there is no 'deeper meaning' to having such unhealthy cravings and having to fight them back. It is no benign, divine set-up in which we can prove the strength of our character by resisting temptation. That is mere moralistic rationalisation of us being uncomfortably stuck in the middle between our genetically inscribed cravings and the culturally produced overabundance. How should we solve this type of mismatch situations? Contrary to what is sometimes assumed, there is no principled reason why we should prefer *changing our cultural surroundings* to *changing our biological determinants.* As Buchanan (2008, pp. 16-18) notes: if for instance some environmental change would cause a basic foodstuff to become partly toxic, we could then either manipulate some aspect of our environment to extract the toxins from the food, or we could manipulate some aspect of our digestive tract to make us resistant to the toxin. The basic heuristic is simple: *whatever works best,* all externalities taken into account.

Thirdly, and most crucially, just as we are physically mismatched to our new surroundings in many respects, so are we *mentally* mismatched to that environment we rightfully call our true home: the *moral* world. No matter what ethical theory one subscribes to or whether or not one accepts some of the stronger claims of evolutionary psychology, it cannot be denied that humans usually strive to live by moral rules yet frequently experience countless problems along the way which make them fall short of the moral goals they aim for. Lack of moral insight, lack of moral resolve and the interference of many immoral desires play crucial roles in sidetracking us from living the way we think we should. Can we simply reiterate our previous conclusion and say that there is no 'deeper meaning' to having such *immoral* cravings and having to conquer them? This is a trickier question.

Many philosophers believe that true morality requires that the moral actor was free to do otherwise. If you were *programmed* to do the right thing (regardless of whether that 'programming' was natural or artificial), then you have not really acted from *moral resolve*. You will then simply have acted from *instinct* or *indoctrination*, they argue, and that instinct or indoctrination may well have

driven you to do some very wrong thing, had you been programmed differently.[101] Interestingly, even one of the most outspoken advocates of human enhancement, Harris (2011), has recently critiqued 'moral enhancement', i.e. the inducement of moral behaviour via HET, specifically such moral enhancements that would erase the *"freedom to fall"* required for moral *choice.*

However, in our view such critiques of moral enhancement are based on a faulty framing of the issue, in that they assume that being influenced by a moral enhancement would necessarily be comparable to being influenced by, say, alcohol, which not only alters your behaviour, but also lessens your capacity to critically assess and direct your own feelings and conduct. We see no reason to believe that moral enhancement would necessarily turn us into such a sort of 'moral drunk'. Even if a person would be 'propped up' to feel and behave morally, all other things remaining equal, such a person would nonetheless retain the capacity for *critical self-assessment*, every bit as much as the ordinary, 'natural' person who is constantly being wooed and swayed by all sorts of moods and desires both moral and immoral, yet retains the 'moral superstructure' of critical self-assessment. So, even if a person was induced by some means to no longer experience all that wooing and swaying but instead felt *spontaneously compelled* to do the right thing, it is unwarranted to suppose that she would then also *necessarily lose* her capacity to rationally assess what she is doing and her ability to make a moral choice on what course of action to take. Now, if a *specific variety* of moral enhancements had such a twofold effect of (a) inducing a desire to act morally and (b) obliterating or reducing the capacity of the person in whom such a desire has been induced to critically assess his own feelings and conduct, then to the extent that this second effect is induced, such an 'enhancement' would indeed merit disapproval, for it would lower us to the level of 'moral drunks'.[102]

The only odd, but in my view very welcome difference with our natural predicament, is that a properly morally enhanced person would find little reason to act contrary to her feelings, because those feelings would have become highly attuned to her moral beliefs thanks to her moral enhancement. Experiencing *feelings of immoral temptation* is not a prerequisite for moral choice, no such 'original sin' is needed. It suffices that one can *rationally assess one's own feelings and conduct.* One does not have to *feel torn* between different courses of action, to be able to choose the good and

[101] See Hursthouse (2002) for an interesting take on Kantian perspectives on this issue.

[102] Perhaps this conclusion may, all in all, differ not *in kind* but only *to a degree* with Harris' (2011) argument: skeptically and briefly, he does allow for the theoretical possibility of such 'proper' moral enhancement being conceivable. Nevertheless, Harris (2011) regards that not only as highly improbable, but also as something different in kind from those moral enhancements proposed by the authors he engages with (Douglas, 2008; Perrson & Savulescu, 2008). I hope to have presented here a first sketch of the basic way in which moral enhancement might effectively leave our 'freedom to fall' fully intact, and thus become truly deserving of its name.

resist the bad. One just needs to *be emphatically aware* of there being such different courses of action, and if moral HETs should allow you to be unhindered by immoral inclinations so that you can get a clear, level-headed view of what's right and feel no desires to do something other than the right thing, then that is so much for the better. Hence it may be concluded that there is no 'deeper meaning' in experiencing all sorts of immoral cravings and inclinations in oneself and having to overcome them. If moral enhancement could 'set us straight inside' and make it easy to do the right thing, for the right reason, without undermining our rational capacity to choose otherwise, that would be quite marvellous.

Even if human nature were noble, why not make it nobler still? Some unconventional, but logical religious arguments

Imagine, counterfactually, humans to be neither morally stunted nor morally malformed by their RHSN. Arguably, such a view of human nature is only possible from a religious perspective that believes man to have been created just as he should be by some benign Creator. Even from such a religious perspective on human nature as a proper and good thing to be respected, it is far from obvious why HET should be shunned and our RHSN should be indefinitely conserved in its status quo condition. Even if our life is a God-given gift for which we humbly owe Him eternal thanks, there is no reason why He would not allow us to further ennoble our existence. Basic theistic dogma actually provides a lot of leeway for accommodating the liberty, perhaps even the *religious duty* to enhance ourselves. As Parens (2009) points out, no lesser source than the Book of Genesis contains a celebration of genetic engineering:

> Jacob, the very one whose name would become Israel, was "the first genetic engineer"; he was the one with the creativity to fashion a device ("rods of poplar and almond, into which he peeled white streaks" (Gen. 30:38) with which he induced his uncle's goats to produce only the valuable "speckled and spotted" (Gen. 30:39) young. According to Genesis, and it seems to me much of Judaism, our responsibility is not merely to be grateful and remember that we are not the creators of the whole. It is also our responsibility to use our creativity and mend and transform ourselves and the world (Parens, 2009, p. 189).

Indeed, Judaism certainly seems open to such interpretations. Rabbi Azriel Rosenfeld, for instance, states the following in his article '*Judaism and gene design*': *"Our sages recognize, and perhaps even encourage, the use of prenatal (or better, preconceptual) influences to improve one's offspring"* (Rosenfeld, 1972, p. 75) Mark Sagoff, in his contribution to the book '*Is human nature*

obsolete?' (in Baillie & Casey, 2005) mentions examples of openness to the idea of human enhancement not only from Judaism, but also from Catholicism[103] and Protestantism[104]. Noteworthy is also the Mormon Transhumanist Association's statement that: "*Mormonism and Transhumanism advocate remarkably similar views of human nature and its future: material beings organized according to law, rapidly advancing knowledge and power, imminent fundamental changes to anatomy and environment, and eventual transcendence of present limitations"*.

Indeed, material human enhancement may even be seen as precisely the faithful behaviour God expects from his flock, and the research and development of HETs then becomes the divinely inspired path to life in the likeness and service of God. The freedom we have to change our own nature, for better or worse, can be perceived as part of the *moral mission* God has given to mankind. In labouring to properly enhance ourselves, we prove ourselves truly thankful of his gift, a gift that we must not squander by being slothful. To lazily 'let Nature take its course' would be to commit the sin of sloth, if not also the sin of succumbing to heathenish nature cultism, and pious virtue would lay in the diligent effort to make the 'seeds He has sown in us' come to full bloom (see Mill, 1904[105]). In this view, God may perhaps only take pity on those repenting sinners who never waver in their labour to amend and improve themselves, who are never content with their crooked selves and the job they have done so far at stewarding the natural world, and who thank God for being so good to grant them access to the means to redeem themselves.

Such a religious narrative would echo the classic of Giovanni Pico della Mirandola (1999), '*Oratio on the dignity of man'* – the hallmark text that helped introduce a humanistic brand of Christianity. In that *Oratio*, della Mirandola lets the (allegorically plural) Creators address Adam as follows:

> The nature of all other creatures is defined and restricted within laws which We have laid down; you, by contrast, impeded by no such restrictions, may, by your own free will, to whose custody We have assigned you, trace for yourself the lineaments of your own nature.…We have made you a

[103] Citing theologian Bernard Haring in saying that humanity may "freely interfere with and manipulate the function of his bios (biological life) and psyche insofar as this does not degrade him or diminish his or his fellowmen's dignity and freedom" (in Sagoff, 2005, p. 84).

[104] Paraphrasing theologian Ronald Cole-Turner's cautious approval of genetic manipulation to improve the conditions of life (Sagoff, 2005).

[105] Mill (1904) pursues an even more radical theological line of thought, in which Nature can be seen as something depraved, bordering the demonic: something man must amend or overcome, so that we may reconnect with the divine.

> creature neither of heaven nor of earth, neither mortal nor immortal, in order that you may, as the free and proud shaper of your own being, fashion yourself in the form you may prefer. It will be in your power to descend to the lower, brutish forms of life; you will be able, through your own decision, to rise again to the superior orders whose life is divine (della Mirandola, 1999 [1486]).

Whatever the ultimate value of the preceding deductions from theist dogma, it is clear that a multitude of religious narratives would support or even mandate the proper use of HETs.

Default biological causation may be more alien to us than intentional artificial causation

Many authors condemn the idea of having artefacts exert an internal influence on our personal, volitional acts, while accepting as something neutral or even positive the ways in which natural determinants exert the very same sort of influence on us. In the following paragraphs, I will defend the opposite view.

Consider how Sandel (2007) judges the difference:

> One aspect of our humanity that might be threatened by enhancement and genetic engineering is our capacity to act freely, for ourselves, by our own efforts, and to consider ourselves responsible – worthy of praise of blame – for the things we do and for the way we are. It is one thing to hit seventy home runs as a result of disciplined training and effort, and something else, something less, to hit them with the help of steroids or genetically enhanced muscles…As the role of the enhancement increases, our admiration for the achievement fades (Sandel, 2007, p. 25).

Sandel (2007) seems to conflate two unrelated levels of analysis. He compares the following situations:

1. Player A: hits seventy home runs based on disciplined training and effort;
2. Player B: hits seventy home runs based on steroids or genetically enhanced muscles.

The glitch, I submit, lays in the conflation of two different functions which the enhanced muscles can be thought to perform.

In Sandel's (2007) view, the enhanced muscles *substitute for the disciplined training and effort*. The result is that disciplined training and effort, the locus of active engagement of the baseball player, *becomes redundant*. Therefore, the enhanced sporting performance becomes a largely *passive* result of the enhanced muscles. But if this is the problem, then this is not an argument about enhancement at all, but about *effortlessness*. For we may just as well imagine a Player C, who has such a *natural* muscle mass, that he too can hit those seventy home runs without any training or effort – a great 'natural talent'. Disciplined training and effort would be just as redundant for this natural Player C as it is for the artificially enhanced Player B. But now imagine that Player B and Player C would no longer content themselves with beating Player A without breaking a sweat, but would instead try to play their best game, against each other. Then both the 'enhanced' (B) and the 'natural' (C) player would invest in disciplined training and effort once again. Or they could find a way – again, natural or unnatural – to

further raise the threshold of effortlessness. But that would automatically create a new peak performance that can be reached only when effort is added to what can be achieved effortlessly. In that upward dynamic, the locus of active engagement moves upward too, allowing us to continue to hold both players responsible for their peak performance which will always entail adding the surplus of effort, so that praise can again be showered over both of them.

So the enhanced muscles in fact do *not necessarily* substitute for the disciplined training and effort, as Sandel (2007) frames it, but instead may turn out to *substitute for the natural muscles.* The result is that disciplined training and effort, the locus of active engagement of the baseball player, does *not necessarily* become redundant for the enhanced Player B, but instead *may begin at a new, elevated baseline*. And so Sandel's (2007) moral conclusion, which is to criticise in Player B the supposed lack of what is lauded in Player A, namely his *effort*, does not automatically apply. Player B may well apply his effort to perform even stronger acts of sportsmanship, and in the event that he would perform effortlessly, that would not be because of his enhancement itself, but because of his laziness and indulgence.[106] Artificial enhancement and the loss of effort or active agency are not intrinsically connected. Some enhancements may have that effect, but others may not:

[106] Sure enough, *in common practice* HETs may well be mostly used as 'easy way out', i.e. as a short-cut to perform things that otherwise would require true effort on their behalf. Such 'easy riders' would then content themselves with having effortlessly done all that is normally required of them and pay no heed to the moral call to really apply themselves in order to fulfill their full potential – a potential that will be heightened even more as they obtain access to HETs. As such, those easy performances would indeed only superficially resemble actual *achievements:* performances that can be accounted to the person who performed them. These hollowed performances then become nothing to laud that person for, and to the extent that such 'easy riders' would still want to extract such approval from others, their short-cuts should indeed be unmasked and dismissed as 'cheap tricks'. However, it should be stressed that even though such short-cut behavior might quite possibly be a temptation that many and perhaps even most people would succumb to and that the availability of HETs might add new, hard-to-resist temptations to indulge in such undignified short-cut behavior, these are no intrinsic objections to HETs. This is a distinction Kass (2002), for instance, fails to make in his *categorical* dismissal of all human enhancement: *"Homogenization, mediocrity, pacification, drug-induced contentment, debasement of taste, souls without loves and longings – these are the inevitable results of making the essence of human nature the last project of technical mastery. In his moment of triumph, Promethean man will become a contented cow."* (Kass, 2002, p. 48). We can share with Kass (2002) the conviction that these are indeed the many *practical* dangers of offering HETs to persons who are likely to be slothful and deceitful, as many humans probably are. Nevertheless, adequately virtuous humans will forego such 'cheap tricks', and will be able to use HETs to, as Bostrom (2008) puts it, *"increase our zest for life, infuse us with energy and initiative, and heighten our capacity for love, desire, and ambition"* (p. 189).

"*[E]nhancements would not transform us into passive, complacent, loveless, and longing-less blobs. On the contrary, they could increase our zest for life, infuse us with energy and initiative, and heighten our capacity for love, desire, and ambition"* (Bostrom, 2008, p. 189).

But if enhancement and effortlessness are not intrinsically linked, then the only remaining differentiation is this:

1. Players A and C play baseball, starting with naturally grown muscles;
2. Player B plays baseball, starting with artificially grown muscles.

But where can the moral difference lay between naturalness and artificiality, if they are – in principle[107] – equivalent with respect to the possibility of effortlessness?

To make some sense of the argument, we have to focus on the fear that caused Sandel (2007) to ring the alarm. There is something in HETs which makes him fear that we may lose *"our capacity to act freely, for ourselves, by our own efforts, and to consider ourselves responsible"* (Sandel, 2007, p. 25). This may be brought about by the fact that some HETs involve the insertion of external, material things into the body, so as to exert an internal influence on our personal, volitional acts. Such 'invasive procedures' disrupt the intuitive way we think about how all our intentional activity comes about, namely via the directives of our personal will. As soon as something other than our personal will (co)causes our actions, we start feeling somewhat like a puppet – someone who is being played upon by someone or something else, a person who is not entirely herself. We start losing our sense that we 'act freely', that our acts are caused by 'our own efforts', and that it is us, our 'selves' we can hold responsible' for those actions.

Admittedly, HETs *acutely* disrupt this intuitive self-understanding. But we have to go full circle here. For the intuitive self-understanding of our 'free self' is premised on the dualistic assumption that our 'self' somehow escapes all material causation, an assumption which has been proven totally mistaken by all we have learnt so far about the way the brain developed and how it operates.

Indeed, even though we spontaneously slide into the dualistic delusion as soon as we stop consciously reflecting on our scientific self-understanding, we know that in fact, we are not our own 'prime mover', we are not a *causa sui*, but rather our mental life is submerged in universal material chains of cause and effect. As Pinker (2003) recalls the debasing effect of accepting the findings of neuroscience:

> One can say that the information-processing activity of the brain *causes* the mind, or one can say that it *is* the mind, but in either case the evidence is overwhelming that every

[107] But again: perhaps not at all in common practice.

> aspect of our mental lives depends entirely on physiological events in the tissues of the brain (Pinker, 2003, p. 41).

When this awareness of our own material determination washes over us, it can completely overhaul our intuitive dualistic delusion and leave us warped and deranged, believing all free will and all personal responsibility is illusionary – we suddenly feel like true puppets on a string, completely overpowered, and this self-understanding is only aggravated by the fact that there is no well-meaning puppet master that 'plays us', only pointless evolutionary dynamics. Our imagined wall between the free, seemingly immaterial mental world and the causally chained material world crumbles and what goes down with it is the illusion that many people consider to be the cornerstone of ethical thought, the belief that the mind is *"a control panel with gauges and levers operated by a user – the self, the soul, the ghost, the person, the 'me'"* (Pinker, 2002, p. 42).

Fortunately, such a disruptive self-awareness only washes over us in the very rare moments where we are forced to reflect intensively on the true, causally and biologically determined nature of our existence – such as in philosophy classes, science classes and, every once in a while, courtrooms where a desperate 'innocence on account of compulsion' defence is being pleaded. Most of the time, however, we can leave our dualistic delusion unchallenged: even if we are philosophically aware of its delusional nature, in everyday practice we enjoy the delusion in all tranquillity.

That cherished tranquillity in everyday existence is now threatened by the possibility of human enhancement, bringing the disruption of the dualistic delusion up close and personal. For here we do not just have to reckon with a reflective insight without direct practical repercussions. Instead, we now face practical, material disruptions of our dualistic delusion in our daily lives, as some HET may demonstrate before our very eyes how material interventions can modulate, overtake and transform our very moods, feelings, desires and intentions. But how is such material determination of our mental life by HETs any different than the everyday material determination of our mind by the default processes of neurological causation? It is estrangement either way.

Failure to acknowledge this may in part be due to the operation of the self-defence mechanism of 'cognitive dissonance', whereby highly undesirable factual insights are 'pushed away' by emotionally charged irrationalities. A passionate but ultimately unsubstantiated belief in a supposed duty to 'be natural' may serve this purpose of pushing away disturbing facts about the material foundations of our mental life and us being the product of pointless, pitiless evolution.

In this respect, HETs force us in the following lose-lose situation. We cherish our dualistic delusion, but we are now forced to give it up and must choose between two types of material determinism. Either we prefer the default, natural determination of our self, or we prefer the interventionist, 'enhancing' determination of our self. The only rational choice would seem to be the second type of determinism, because it is in fact *less* alienating to us in two important respects. Firstly, it does not force us to defend what is virtually indefensible, namely that our 'stunted, crooked' and existentially absurd evolutionary endowment is somehow optimal and sacrosanct for us. And secondly, by applying HETs on ourselves, we achieve a partial (albeit quite quaint) liberation from

material determinism after all. We become a sort of 'ultimate feedback loop' to ourselves: out of our own deliberations on what is best for us, we can begin to extract ourselves out of the stunted, crooked RHSN that emerged pointlessly out of hominid evolution, and in the closest material approximation of 'pure freedom', choose our own determinants:

> Homo sapiens, the first truly free species, is about to decommission natural selection, the force that made us. There is no genetic destiny outside our free will...Soon we must look deep within ourselves and decide what we wish to become (Wilson, 1999, p. 302).

The prospect of being able to control, modulate and change those *'tissues of the brain'* thus seems to bring back with a vengeance part of the old dualistic dream of being *"a self, a soul, a person, a 'me'"* (Pinker, 2002, p. 42) able to steer the body and shape our own thoughts by *"operating on it via a control panel with gauges and levers"* (Pinker, 2002, p. 42). Thus turning ourselves into beings that can emerge (theoretically) wholly out of their own volition, we in fact revindicate the belief that humans are indeed beings that are fundamentally 'free', in possession of a will that can somehow escape its own material determination – if not causal determination *in abstracto*, at least we escape the *biological* determination that lock all other life forms in certain fixed bodily and behavioural patterns. Returning to della Mirandola (1999): we are, by our very nature, beings who are at liberty to choose for ourselves our own nature, a self-awareness made even more explicit by Jean-Paul Sartre some fifty years ago. What HET add to this 'traditional' existentialist self-understanding[108], is that it now turns out that this is *materially* the case and, to the extent that a number of specific HETs is already available, that this has become *a pressing practical problem* for every ordinary person. To deny that self-awareness is at best immature, at worst cowardly:

> [Man] cannot find anything to depend upon either within or without himself. [H]e is without excuse. [O]ne will never be able to explain one's action by reference to a given and specific human nature. [M]an is condemned to be free (Sartre, 1964, p. 31, translated from French).

[108] The philosophy of Jean-Paul Sartre, of which a key conclusion is quoted here, was premised on metaphysical notions that cannot be squared with contemporary scientific insights in the biological underpinnings of human behavior and decision making. Therefore, the quote used here should not be seen as an acceptation of Sartre's untenable metaphysical belief that the free will of humans somehow manages to escape all material causation. On the contrary, I want to point out the irony that the 'traditional' existentialist conclusion that 'man cannot find anything to depend upon either within or without himself' gains a new relevance precisely because *wholly material means* now seem to confront us with such a fundamental indeterminacy.

In conclusion, only with a *mauvaise foi* or false conscience, could we hold that it is somehow more authentic that a person is predetermined by 'natural' factors given by morally meaningless evolutionary processes, than by factors he himself chose to insert in himself. Contra Sandel (2007), in this sense it is indeed something else, but something *more*, to fully take charge over your own determinants in pursuing epic athleticism (and the same holds for all virtuous pursuits), as long as the enhanced athlete continues to add to that strengthened body the intentional dedication, effort and focus we rightfully consider central to the athletic ethic.

Enhancement as redress of natural unfairness

In the previous section, Sandel's (2007) objection regarding *effort and fairness* had to be untangled from the underlying objection of *'feeling overtaken'* by artificial means and thus losing *accountability* for one's 'doped performance'. Surprisingly, the objection of being overtaken turned out to be even more the case for everyday *un*enhanced performance (which is a deeply biologically predetermined affair), than it was for enhanced performance (which in principle alleviates our everyday 'biological passivity' by making our biological predeterminations something we can partly shape ourselves via HETs). In the following paragraphs, I deal with another 'effort-fairness' objection Sandel (2007) raises, namely that enhancement represents a form of *'hyperagency'* that erodes our sense of undeserved giftedness, which he considers to be a fundamental basis for solidarity:

> Why do the successful owe anything to the least-advantaged members of society? The best answer to this question leans heavily on the notion of giftedness. The natural talents that enable the successful to flourish are not their own doing but, rather, their good fortune – a result of the genetic lottery. [Thus] it is a mistake and a conceit to assume that we are entitled to the full measure of the bounty they reap[.] We therefore have an obligation to share this bounty with those who, through no fault of their own, lack comparable gifts (Sandel, 2009, p. 87)

I fully agree with the abstract moral argument underlying this quotation: if one has obtained decisive advantages as a matter of brute luck, through no effort on one's own behalf, a basic moral duty impels such an undeservedly 'gifted' person to share his luck with those who are faultlessly unlucky.[109] However, there seems to be a second glitch in Sandel's (2007) reasoning, yielding problematic consequences.

[109] The extent to which this is the case is a persistent bone of contention between 'liberals' and 'socialists' that we do not have to get in to too deeply here. It may suffice to say that on one end, if one chances upon a 50 euro bank note in an empty street, one is not necessarily duty-bound to give this money to some charity but can

In Sandel's (2007) view, we are obliged to share 'the bounty' or *'the fruits of'* [110] our good fortune. But why stop there? For what if *the good fortune itself* can become a matter of redress? If HETs could give the naturally disadvantaged access to what others effortlessly and unfairly gained naturally, surely they should be allowed to 'level the natural playing field' that was so wantonly unlevelled by the random distribution of natural gifts via the 'natural lottery'? This question is reminiscent of Rawls' (1971) 'original position', and in fact it is something Rawls (1971) himself already hinted at in *A Theory of Justice:*

> It is also in the interest of each to have greater natural assets. This enables him to pursue a preferred plan of life. In the original position, then, the parties want to insure for their descendants the best genetic endowment (assuming their own to be fixed). The pursuit of reasonable policies in this regard is something that earlier generations owe to later ones, this being a question that arises between generations. Thus over time a society is to take steps *at least* to preserve the general level of natural abilities and to prevent the diffusion of serious defects (Rawls, 1971, p. 108)[111]

So why insist on redressing unfair advantage only at the time of reaping, when HETs may put us at liberty to redress them at the time of sowing?

There is something inconsistent in wilfully abstaining from intervening in a process of unfair, random distribution, only to then acknowledge the unfairness later on, and then feeling obliged to

---

pocket it without becoming morally stained; on the other end, if one was so lucky to be born into a happy, affluent family, spontaneously grows up to become cheerful, handsome, creative, intelligent and have the good fortune to succeed in every venture undertaken, in such a case of extremely beneficial giftedness one may indeed feel duty-bound to 'give back to the community', out of which one has rather undeservedly risen to the highest ranks of happiness and success. So at *some* point and to *some* extent, undeserved giftedness brings about a moral duty to 'share the treasure', and that is the only moral conclusion we require for our current argument.

[110] I feel obliged to point out the fact that in her article *'What is and what isn't wrong with enhancement?'*, Kamm (2009) misquotes Sandel as having written in his 2004 article *'The case against perfection': "A lively sense… that none of us is wholly responsible for his or her success makes us willing to share the fruits of our talents with the less successful"* (Sandel, 2007, as misquoted in Kamm, 2009, p. 93) From 'makes' onwards, that phrase is nowhere to be found in Sandel's article, which is in fact reprinted in the same book that holds Kamm's article (Savulescu & Bostrom, 2008). Luckily, Kamm (2009) does not seem to misrepresent Sandel's argument by this misquotation, but nevertheless rephrases it in a way that makes it convenient to take issue with.

[111] Note that the 'at least' seems to allow for further development of the original position argument to include the possibility of human enhancement beyond the maintenance of a normal level of health and the prevention of diffusion of serious defects. When taking into account Rawls' (1971) concept of primary goods, it may however turn out that procuring enhancement would only be a marginal, perhaps even negligible, redistributive concern.

redress that random distribution *after the initial damage has been done*, via the practice of the naturally privileged helping out the naturally underprivileged. Such an argument smacks of paternalist aristocracy, where, out of a half-baked moral awareness that *"the natural talents that enable the successful to flourish are not their own doing but, rather, their good fortune"* (Sandel, 2009, p. 87), a *natural* 'genocracy' or 'genobility' (Mehlman, 2007) give a measure of charity to the naturally underprivileged, but at the same time *categorically denies the underprivileged access to the privilege itself*.[112]

True enough, HETs could be used not only to redress natural unfairness, but for the exact opposite: to aggravate existing disparities in privilege, perhaps even to the extent of creating a 'run-away effect' wherein the wealthy of today enhance themselves while the poor get left behind in their stunted old RHSN, resulting in a fundamental split-up of society into an *enhanced* 'genocracy' insulating itself *biologically* from the backwardly natural 'gen-paupers' (a gloomy prospect often feared, see for instance Annas, Andrews, & Isasi (2002), Silver (2006) or Mehlman (2007)). The acceptance of HETs can thus theoretically 'swing both ways', either used as a tool to redress of what is perhaps the prime source of inequality, i.e. the natural lottery, or instead as a tool to strike the ultimate blow in separating the haves from the have-nots.[113]

The moral and legal enforcement of the natural lottery as a proper institution for distributing capabilities in the population, however, *swings only one way:* it forces the populace under a

[112] Turning to moral psychology, we can point to the resemblance to unsavory logic of the fortunate who would surely gladly *help* those less fortunate, but are scandalized by the thought of truly *sharing their fortune*, for of course, such true benevolence would make them lose their exalted position of being 'the naturally blessed', and also make them lose the moral vanity of being 'the philanthropist hero': the one that gives alms to the needy, in all her kindness.

[113] Weighing the probabilities here would require firing up a sociological, political debate, which falls well beyond the scope of this paper. It seems starry-eyed to imagine it more likely that HETs will be used for 'enhanced equality'. However, certain HETs, among which a cognitive enhancer that improves executive function, seem to have a greater enhancing effect on people with lower-than-average capacities (Farah et al., 2004). Such findings could facilitate a policy to use HETs for the emancipation of the less-endowed. Another argument for such a policy goes that the equal or equalizing administration of HETs may prove to be a much more feasible way to aid the less fortunate: *"In comparison with other forms of enhancement that contribute to gaps in socioeconomic achievement, from good nutrition to high-quality schools, neurocognitive enhancement could prove easier to distribute equitably"* (Farah et al., 2004, p. 423). All in all, these are only meager findings, not at all capable of ensuring that HETs will be used to 'enhance equality'. To ensure that the introduction of HETs swings the way of greater equality, perhaps nothing would be more potent than an *active empowerment of the less-endowed to stake their claim on the HETs,* but that is, of course, a very tall order.

distributive regime where many end up not only deeply underprivileged, but also stripped from the right to fundamentally redress their natural incapacitation should HETs ever offer them the possibility to do so.[114] To declare the natural lottery *morally and perhaps even legally binding* save for the caveat of curing severe illness, as many bioconservative authors advocate (Annas, Andrews, & Isasi, 2002; Fukuyama, 2002; Kass, 1997, 2003; President's Council on Bioethics, 2003), thus seems highly contentious, as it may come across as turning a natural unfairness into a political injustice, thwarting fundamental rights.[115] As effective HETs become available, such prohibitions might provoke moral outrage amongst those who see no reason to submit themselves and their children to the lottery, and instead wish to provide themselves with the best life possible (Savulescu, 2001; Harris, 2007).[116] It is misguided to combat the scenario of an 'enhanced genocracy' with an enforcement of the natural lottery. Not only is that process inherently absurd, to wilfully conserve it when proper means have become available to allow the moral steering of those processes, is to enact a 'natural genocracy'. On this general level of analysis, the only proper way to combat the 'enhanced genocracy' scenario is to utilise the HETs in the 'third way' politics that can be dubbed 'enhanced equality'.[117]

---

[114] As an anonymous reviewer rightly pointed out, the genetic lottery is very complex and multi-faceted, and although some people come out better than others, they very rarely are genetically more advantaged throughout. Indeed, every person that can be thought of as having drawn a ticket in the natural lottery that yields *more* optimal determinants than average, will still have *some* suboptimal determinants in his 'genetic bundle' as well. Nevertheless, we can acknowledge this 'mixed genetic blessing' argument (as well as a possible further objection of there being a fundamental *incommensurability* of different human value systems), and still argue that there are clear cases abound of people being, *on the whole*, born extremely to somewhat lucky and others born somewhat to extremely unlucky. And so it seems possible to retain the idea that the natural lottery does swing in only one way, but we must add: it swings *with less force* in the direction of a 'natural genocracy' than a divisive policy of 'enhanced genocracy' might.

[115] Of course, most bioconservative authors are deeply compassionate with the sick or otherwise incapacitated, and propose to soften these harms through other means, such as proper professional and societal care.

[116] Therefore, it seems virtually impossible that pluralistic democracies could ever successfully sustain such categorical bans.

[117] When analyzed to a further degree of (crucial) detail, such a policy of 'enhanced equality' is itself prone to many glaring dilemmas, if not outright paradoxes. For in its most superficial form, it would lead to the totalitarian *reductio ad absurdum* of having a state duty-bound to create 'biologically uniform citizens'. Nevertheless, such objections certainly do not discredit the basic insight that in the basic threefold set-up of 'enhanced equality'-'enhanced genocracy'-'natural genocracy', the last two are fundamentally unethical, and must be dismissed at root. It is as clear however, that within the very general 'enhanced equality' category, many equally outrageous and abhorrent policies have to be combated, and only after substantial further qualification could an 'enhanced

In light of these arguments, it would be most interesting to see how theories of social justice such as Sandel's (2007) can be clearly differentiated from the unsavoury position that by giving redistributive alms, the naturally privileged are redeemed from deeper moral qualms. Quite quizzingly, Sandel (2007) does conclude his article and his book (an extension of that article) '*The case against perfection'*, with a mention of Robert L. Sinsheimer who hints at objections similar to the ones raised here in his 1969 article '*The prospect of designed genetic change'*. Sandel (2007) notes how Sinsheimer wrote hopefully of rescuing *"the losers in that chromosomal lottery that so firmly channels our human destinies"* (Sinsheimer in Sandel, 2009, p. 88; Sinsheimer in Sandel, 2007, p. 97). However, he does not seriously engage with the argument. Therefore, I have made the argument for 'enhanced equality' more substantial here, and made the possible (but quite possibly misguided) unsavoury interpretation of Sandel's (2007) reasoning more explicit.

equality' policy become a truly coherent and ethical whole. I recommend '*From chance to choice – Genetics and justice'* (Buchanan, Brock, Daniels, & Wikler, 2000) as a *locus classicus* for this further debate, issues which are expanded upon in Buchanan's new book '*Beyond humanity?'* (2011).

Conclusion: Shouldn't we be natural *at all?* Seeing humans as Ships of Theseus

> Laziness and cowardice are the reasons why so great a proportion of men, long after nature has released them from alien guidance, nonetheless gladly remain in lifelong immaturity (Kant, *What Is Enlightenment?).*

In this paper, I have pursued five fundamental challenges to the idea that 'we ought to be natural', understood as the moral imperative that we ought to live by our human nature, more specifically our 'recalcitrant *homo sapiens* nature'. Such an imperative is often advanced in the ethical and legal debates surrounding human enhancement technologies, where it is often dismissed, sometimes out of hand, by applying the apparent knock-down argument that is 'the naturalistic fallacy':

> [There are] those who argue for a distinct essence, a kind of template of humanity that somehow is in there as a core that cannot be touched or changed or manipulated without loss of who we are – they are nervous conservatives who worry that the bearings will be lost if we admit that what we are is a jumbled set of mishmash traits evolved and designed to handle a random environment from the past that we don't have to care about any more. The antimeliorists are making the conceptual error, that the way we are is the way we should be. I'm submitting that what we know from evolution, from Darwin's day on, is that the way we are is an interesting accident. And it tells us certain things about what will make us function well, but it doesn't tell us anything about the way we should be or what we should become or how we should decide to change ourselves (Caplan, 2006, p. 38).

Potent as that basic argument may be, it clearly does not convince many authors who remain avid, principled 'antimeliorists' or 'bioconservatives'. Against this backdrop and to enliven the entrenched debate, I have attempted to marshal the following more or less off-beat arguments as to why we should not consider ourselves morally obliged to conserve our RHSN, thus 'flogging the dead horse of nature's normativity, with a twist':

1. As we undertake open-ended worthy pursuits such as the search for truth and physical prowess, aspects of our RHSN will at some point come to hinder those pursuits, and eventually arrest progression further down the line.
2. If we analyze the inner traits of our RHSN, we discover numerous trappings in it that require alteration if we wish to succeed in living the way we morally want to.
3. Even if our RHSN were a noble thing and/or a benign creation, a gift for which we owe humble thanks, that idea alone does not imply that we should maintain the natural status quo. Quite possibly, we might be mandated to not squander that gift, and make something more of it according to a higher moral plan.

4. As 'invasive procedures', HETs acutely confront us with the extent to which our mental life and our sense of identity emerge out of material processes. This deeply challenges the 'dualistic delusion' that is central to our conventional self-understanding as 'free agents, worthy of praise or blame'. However, HETs nevertheless make us more free and more responsible for ourselves: they allow us to choose for ourselves some of the material determinants that make up who we are, thus to some extent freeing us from emerging out of material determinants that we previously had to accept in full passivity.[118]

5. HETs can be thought of as means to gain unfair advantage, and more generally, allowing HETs may come to erode the sense of 'undeserved giftedness' which Sandel (2007) considers to be the basis of solidarity of the natural haves for the natural have-nots. However, HETs may (perhaps improbably, but still) be put to an opposite use as well, to 'level the natural playing field'. On the contrary, declaring the 'natural lottery' morally or legally binding seems to set the 'brute luck at birth' in stone, denying the naturally incapacitated to gain through HETs what they were gainsaid by nature.

Should we then simply *do away with human nature altogether?* No, not at all. The given arguments only serve to dispel the belief in there being an *intrinsic* value encapsulated somewhere in

[118] A promising follow-up argument I will examine in further research is this: perhaps we do not *want* to have to consider ourselves to be so deeply responsible for who we are. Such a desire for passivity may explain some of the bioconservative resistance to HETs. For instance, Sandel (2007) candidly admits that *"the real problem is the explosion, not the erosion, of responsibility. As humility gives way, responsibility expands to daunting proportions. We attribute less to chance and more to choice…One of the blessings of seeing ourselves as creatures of nature, God, or fortune is that we are not wholly responsible for the way we are"* (p. 87). But there are good reasons to assume that such 'ruthless' freedom seems inescapable: if we choose to continue letting 'nature take its spontaneous course', it will be *us choosing so*. The 'spontaneous', 'natural' process will thus necessarily pass through a human decision procedure. So it seems that *"paradoxically, nature brought within human control is no longer nature"* (Buchanan et al., 2000, p. 84). Expanding this argument may have profound repercussions for, inter alia, the question whether it is at all possible *not to* make designer babies in a world where HETs are readily available, considering that the 'natural babies' will just as much be the outcome of their parents' or their state's *choices*, in casu 'designing' them as 'natural'. In any case it would seem that children, when coming of age, will confront their parents with their reproductive responsibility *either way*. They may ask: 'why did you deliberately enhance me?' Or they may ask: 'why did you deliberately keep me natural?' Chances are they will feel 'designed' on both occasions, a thought that has led Sloterdijk (1999) to (somewhat cryptically) declare every child born in the age of enhancement to be *"convicted to trust"* (p. 7). In a forthcoming paper, I will zoom in on this specific issue of 'responsibility explosion' and the unsettling possibility that this explosion may be inevitable.

our RHSN, *"a core that cannot be touched or changed or manipulated without loss of who we are"*, as Caplan (2006, p. 38) put it. Such a dispelling does not at all imply that we must ban all natural features from our sense of identity and uphold a puritan conception of ourselves as purely abstracted, free and rational minds. True, in part we are such 'free-floating' minds, but we are also nestled deep within a staggeringly intricate *homo sapiens* biological constitution. As a result, we of course seek out specifically *human* experiences, such as family love, friendship, compassion, artistic flights of fantasy, the pursuit of truth, great sex and great food, and we rightfully seek to fill our lives with such experiences that are, in a very broad sense, 'species-typical'. Thus, it is trivially, albeit importantly true that we can inflict harm on ourselves and others if we pursue misguided goals that go too deeply against the grain of our recalcitrant nature. That nature therefore, does have an important *derivative* value (Bayertz, 2003; Singer, 2000).

That being said, our RHSN is no *"seamless web [where] severing one fiber is likely to result in the whole thing unravelling"* (Buchanan, 2009, p. 147). Such a 'high-strung', categorically conservative stance would require *"empirical evidence, not armchair speculation"* (Buchanan, 2009, p. 147) to discount each specific enhancement scenario, which more often than not, is not readily provided by bioconservative authors. Such knee-jerk conservatism cannot be taken at face value, as the historical record shows that many important innovations that are now completely integrated in our 'natural' way of life have provoked hysteria and fears of debasement at the time of their introduction – from vaccination to organ transplantation, from contraceptives to in vitro fertilisation (IVF).

Innovation frequently involves enormous serendipity. As technologies come to influence our existence ever more profoundly, we should continuously re-assess whether they, on the whole, really do serve us right, and do not impoverish our lives more than they enrich them – a technological 'art of living' should be on our minds daily in our plane-flying, internet-surfing, pill-popping lives. Arguably, as an ethical rule of thumb the 'invasive procedures' of HET merit extra caution, not in the least to counterbalance the impatient 'extra enthusiasm' of the transhumanists, brought on by the promise of HETs to finally *"debestialize humanity"* (Sloterdijk, 1999, p. 29).

To maintain a basic sense of species identity to carry with us during this shaky but certain 'exodus' out of our troubled *homo sapiens* beginnings, a variation on the identity riddle of the Ship of Theseus may do us justice. Human nature can be thought of as a wooden ship we find ourselves on, sea-worthy so far but certainly improvable. Inventive and industrious as we are, we contrive all sorts of ways to not only replace worn parts, but to put new and improved parts in their place. Our process of improving self-change is open-ended in two distinct senses: in principle we may *replace all* possible parts of our boat, thus becoming deeply self-shaped, and deeply responsible for who we are; and in principle we may also *improve all* possible parts any way we might want to, bound only by the self-imposed restraints that we conserve its coherence so as to keep ourselves afloat, and that we enhance it into the finest, most dignified vessel we are capable of making.

References

Agar, N. (2004). *Liberal eugenics. In defence of human enhancement.* Malden, MA: Blackwell.

Agar, N. (2010). Thoughts about our species' future: themes from humanity's end: Why we should reject radical enhancement. *Journal of Evolution and Technology, 21(2),* 23-31.

Annas, J. B., Andrews, L. B., & Isasi, R. M. (2002). Protecting the endangered human: Toward an international treaty prohibiting cloning and inheritable alterations. *American Journal of Law & Medicine, 28(2&3),* 151-178.

Bailey, R. (2005). *Liberation biology. The scientific and the moral case for the biotech revolution.* New York: Prometheus Books.

Baillie, H. W., & Casey, T. K. (2005). *Is human nature obsolete? Genetics, bioengineering, and the future of the human condition.* Cambridge, MA: MIT Press.

Berthelot, G., Thibault, V., Tafflet, M., Escolano, S., El Helou, N., Jouven, X., Hermine, O., & Toussaint, J.-F. (2008). The citius end: World records progression announces the completion of a brief ultra-physiological quest. *PLoS One*, 3 (2), 1-5.

Bostrom, N. (2008). Dignity and enhancement. In *Human dignity and bioethics.* The President's Council on Bioethics. Retrieved from http://bioethics.georgetown.edu/pcbe/reports/human_dignity/chapter8.html

Brownsword, R. (2009). Regulating human enhancement: Things can only get better? *Law, Innovation and Technology, 1(1),* 125-152.

Buchanan, A. (2008). Enhancement and the ethics of development. *Kennedy Institute of Ethics Journal, 18(1),* 1-34.

Buchanan, A. (2009a). Human nature and enhancement. *Bioethics, 23(3),* 141-150.

Buchanan, A. (2009b). Moral status and human enhancement. *Philosophy & Public Affairs, 37(4),* 346-381.

Buchanan, A. (2011). *Beyond humanity? The ethics of biomedical enhancement.* Oxford: Oxford University Press.

Buchanan, A., Brock, D.W., Daniels, N., & Wikler, D. (2000). *From chance to choice. Genetics & Justice.* Cambridge: Cambridge University Press.

Caplan, A. (2006). Is it wrong to try to improve human nature? In P. Miller & J. Wilsdon (Eds.), *Better humans. The politics of human enhancement and life extension* (31-40). Londen: Demos.

Caplan, A. (2009). Is the perfect the enemy of the good? *Perspectives in Biology and Medicine*, *52(4),* 624-627.

Cherry, M. J. (2009). *The normativity of the natural: Human goods, human virtues, and human flourishing.* New York: Springer.

Dawkins, R. (2006). *The selfish gene: 30th anniversary edition.* Oxford: Oxford University Press.

Dawkins, R. (1996). *The blind watchmaker. Why the evidence of evolution reveals a universe without design.* New York, Lodnon: W.W. Norton & Company.

Della Mirandola, P. (1999). Oratio on the dignity of man. In P. Brians, M. Gallwey, A. Hussain, R. Law & M. Myers (Eds.), *Reading about the world.* Harcourt Brace Custom Publishing.

Dennett, D. (1992). The self as a centre of narrative gravity. In F. Kessel, P. Cole & D. Johnson (Eds.), *Self and consciousness: Multiple perspectives* (103-115). Hillsdale, NJ: Lawrence Erlbaum Associates.

Douglas, T. (2008). Moral enhancement. *Journal of Applied Philosophy*, 25(3), 228-245.

Dworkin, R. (1999). Playing god. *Prospect Magazine,* 41. Retrieved from http://www.prospectmagazine.co.uk/1999/05/playinggod/

Farah, M., Illes, J., Cook-Deegan, R., Gardner, H., Kandel, E., King, P., Parens, E., Sahakian, B., & Wolpe, P. R. (2004). Neurocognitive enhancement: what can we do and what should we do? *Nature, 5 (5),* 421-425.

Fukuyama, F. (2002). *Our posthuman future: Consequences of the biotechnology revolution.* New York: Picador.

Habermas, J. (2003). *The future of human nature.* Cambridge: Polity Press.

Harris, J. (2007). *Enhancing evolution: The ethical case for making better people.* Princeton: Princeton University Press.

Harris, J. (2011). Moral enhancement and freedom. *Bioethics, 25(2),* 102-111.

Horgan, J. (1997). *The end of science: Facing the limits of knowledge in the twilight of the scientific age.* New York: Broadway Books.

Hughes, J. (2004). *Citizen cyborg: Why democratic societies must respond to the redesigned human of the future.* Cambridge, MA: Westview Press.

Hughes, J. (2009). Beyond human nature. In P. Healey & S. Raynor (Eds.), *Unnatural selection: The challenges of engineering tomorrow's people* (51-59). London: Earthscan.

Hursthouse, R. (2002). *On virtue ethics.* Oxford: Oxford University Press.

Kant, E. (Unknown date of translation's publication). An answer to the question: What is enlightenment? Retrieved from http://www.english.upenn.edu/~mgamer/Etexts/kant.html.

Kass, L. R. (1997). The wisdom of repugnance. *New Republic*, 216(22). Retrieved from http://www.catholiceducation.org/articles/medical_ethics/me0006.html.

Kass, L. R. (2003). Ageless bodies, happy souls. *The New Atlantis, 1,* 9-28.

Kayser, B., Magnon, A., & Miah, A. (2007). Current anti-doping policy: a critical appraisal. *BCM Medical Ethics, 8(2),* 1-10.

Kious, B. (2008). Philosophy on steroids: why the anti-doping position could use a little enhancement. *Theoretical Medical Bioethics, 29(4),* 213-34.

Lippi, G., Banfi, G., Favaloro, F. J., Rittweger, J., & Maffulli, N. (2008). Updates on improvement of human athletic performance: focus on world records in athletics. *British Medical Bulletin, 87,* 7-15.

Maher, B. (2008). Poll results: Look who's doping. *Nature, 452*, 674-675.

McKibben, B. (2003). *Enough. Genetic engineering and the end of human nature.* London: Bloomsbury Publishing.

Mehlman, M. (2007). Genetic enhancement: Plan now to act later. *Kennedy Institute of Ethics Journal, 15(1),* 77-82.

Mill, J. S. (1904). On nature. *Nature, the utility of religion and theism*. Retrieved from http://www.lancs.ac.uk/users/philosophy/texts/mill_on.htm.

Montaigne, M. (2004). *Essais. Livre III.* Quadrige – Presses Universitaires de France.

Perrson, I., & Savulescu, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of Applied Philosophy, 25(3),* 162-167.

Pinker, S. (2002). *The blank slate: The modern denial of human nature*. New York: Viking Press – Penguin Group.

Pinker, S. (2003). The designer baby myth. *The Guardian*. Retrieved from http://www.guardian.co.uk/education/2003/jun/05/research.highereducation.

President's Council on Bioethics. (2003). *Beyond therapy. Biotechnology and the pursuit of happiness.* Washington D.C.: The President's Council on Bioethics.

Rozenfeld, A. (1972). Judaism and gene design. *Tradition, 13,* 71-80.

Sandel, M. (2007). *The case against perfection. Ethics in the age of genetic engineering.* Cambridge, MA: Harvard University Press.

Sandel, M. (2009). The case against perfection. What's wrong with designer children, bionic athletes, and genetic engineering. In N. Bostrom & J. Savulescu (Eds.), *Human Enhancement* (71-90). Oxford: Oxford University Press.

Sartre, J.P. (1964). *L'existentialisme est un humanisme.* Les Éditions Nagel.

Savulescu, J. (2001). Procreative Benificience: Why we should select the best children. *Bioethics, 15(5/6),* 413-426.

Silver, L. M. (2006). *Challenging nature. The clash of science and spirituality at the new frontiers of life.* New York: HarperCollins Publishers.

Singer, P. (2000). *A Darwinian left: Politics, evolution, and cooperation.* New Haven: Yale University Press.

Sloterdijk, P. (2000). Regels voor het Mensenpark. In L. Ten Kate (Ed.), *Regels voor het Mensenpark. Kroniek van een debat* (18-50). Amsterdam: Boom.

Talbot, M. (2008). Brain Gain. The underground world of "neuroenhancing" drugs. *The New Yorker*. Retrieved from http://www.newyorker.com/reporting/2009/04/27/090427fa_fact_talbot.

Tallis, R. (2007). Enhancing humanity. *Philosophy Now, 61.* Retrieved from http://www.philosophynow.org/issue61/61tallis.htm.

Tännsjö, T. (2009). Medical enhancement and the ethos of elite sports. In N. Bostrom & J. Savulescu (Eds.), *Human enhancement* (315-326). Oxford: Oxford University Press.

UNESCO (2005). Universal declaration on bioethics and human rights. Retrieved from http://www.eubios.info/udbhr.pdf.

Wilson, E. O. (1999). *Consilience. The unity of knowledge.* New York: Vintage Books.

PART II: LAW, ETHICS AND ROBOTICS

Section A: Foundations of roboethics

Chapter 11
From robots to techno sapiens: Ethics, law, and public policy in the development of robotics and neurotechnologies

Wendell Wallach
Yale University
Yale Interdisciplinary Center for Bioethics
✉ wendell.wallach@yale.edu

Abstract Robots bear similarities to other technologies whose safety and appropriate use has been addressed by existing ethical standards, professional codes, laws, and regulations. But engineers are continually discovering ways to implement new capabilities. The applications for which robots will be used are expanding rapidly. Robots with even limited sensitivity to ethical considerations and the ability to factor those considerations into their choices and actions will open up new markets. However, if they fail to adequately accommodate human laws and values in their behaviour, there will be demands for regulations that limit their use.

Many of the anticipated ethical, legal, and policy challenges arising from the use of robots for new applications can be addressed incrementally. It will, however, be important to also keep an eye on the broader societal impact of introducing robots into the home, the battlefield, and the commerce of daily life. Over the next twenty years, stand-alone robots, and robotic technologies in combination with neurotechnologies and other emerging technologies will contribute to a transformation of human culture. We will be confronted with the difficult challenge of not just monitoring and managing individual technologies that are each developing rapidly, but also the convergence of many technologies.

Keywords robot, artificial agent, machine ethics, emerging technology, decision making

Introduction

The specific ethical and legal challenges posed by robotics must be considered within the context of the broader societal impact of emerging technologies. The public is generally fascinated by new technologies, and perceives technology as an engine of both promise and productivity. But there is also considerable disquiet as to whether we are surrendering the future to a juggernaut of change that will decimate cherished values and institutions. This disquiet is evident in the worldwide prohibition on human cloning, restrictions upon the sale of genetically modified foods in the EU, controversy regarding research using embryonic stem cells in the U.S., and international regulations prohibiting athletes from using human growth hormones and drugs that enhance performance. Technological innovation offers countless rewards, but also poses dangers that are difficult to predict. How will humanity navigate the promise and perils of the bio-tech, info-tech, and nano-tech revolution?

The various fields (genomics, synthetic biology, nanotechnology, information technology and robotics,

regenerative medicine, and neuroscience) that are contributing to this revolution overlap and converge. The overlap and convergence of research in neuroscience and artificial intelligence will be given particular attention in this article.

Computational neuroscience has become an important tool for revealing information processing properties of various structures within the nervous system. Computer simulations provide laboratories for testing various theories about mental activity. Findings in neuroscience inform strategies for developing discrete cognitive capabilities in AI. The computational theory of mind drives hypotheses regarding the likelihood of reproducing human intelligence artificially. In turn, critics of the contention that mental activity can be reduced to its computational components are pressed to sharpen their arguments, as are critics of the hypothesis that human-level intelligence and consciousness can be reproduced artificially. The convergence of robotics and neuroscience will be realised with the development of advanced neuroprosthetics, in the creation of robots with higher-level cognitive capabilities and artificial general intelligence, and with the emergence of a culture of techno sapiens, individuals who utilise information technology and neurotechnologies to enhance their capabilities.

Television required thirteen years to reach an audience of 50,000,000 while the Internet required only four years. Few predicted the speedy growth of the Internet and even faster adoption of smart telephones, two innovative technologies that have transformed behaviour, communications, entertainment, and education. Michael Polanyi and Karl Popper, two 20th century philosophers of science, recognised that scientific progress can be unpredictable. This is often presumed to mean that scientific development is also ungovernable. Indeed, Popper (1945) and Polanyi (1951) were also both concerned with the dangers posed by any governmental attempts to direct the development of science and restrict the freedom of scientific enquiry.

Nevertheless, the desire to maximise the societal benefits derived from science has always been weighted against the need to minimise harms. Governments are heavily involved in directing scientific development in the form of capital investment and through regulations and regulatory oversight directed at public health, human subjects research, the oversight of animals used in research, the safety of goods and services, the safety of laboratory workers, and, more recently, biosecurity. Criminal and tort law, insurance regulations, professional codes of conduct, guidelines for laboratory practices and procedures, and other strategies for soft governance contribute to a relatively robust system of protections. Deriving benefits from research in genomics and nanotechnology while protecting the public against harms caused by exposure to toxic nanoparticles, pathogenic organisms, or potentially dangerous genetically modified foods has received particular attention over the past decade. Addressing new challenges is largely a piecemeal process of adding new laws and regulatory agencies as needed. Existing policy mechanisms can also be modified.

Debate is underway, for example, regarding whether regulations on research ethics should be lowered to allow field-testing of GM plants for medical applications (growing inexpensive antibiotics), or raised to prohibit animal enhancements or research on synthetic biology.[119] There are, however, serious questions as to whether the cumulative impact of emerging technologies will overwhelm the piecemeal, incremental approach to monitoring and managing their development.

This article will be an exercise in foresight, planning, and anticipatory governance. The unpredictable course of technological development should not be interpreted as meaning that planning is futile. Social theorists such as Arizona State University's David Guston challenge the assumption that unpredictable should be equated with ungovernable. Anticipatory knowledge, including predictable trends, can result from analysis of how specific technologies are likely to be used. In turn, this knowledge can contribute to formulating public policy. Given the potential of emerging technologies to cause considerable harms there is a serious need to develop methodologies for anticipatory governance (Guston, 2010).

But it is probably naive to expect legislators to act upon anticipatory knowledge. Unfortunately, legislative attention to technological concerns is commonly held in abeyance until forced by unanticipated disasters. From thalidomide babies to the Chernobyl meltdown, technology has been complicit in crisis after crisis. Most recently, information technology (IT) played a significant role in the derivative crisis and the 'flash crash.' While not caused by IT, the BP oil spill, like the Challenger disaster, is an example of a crisis resulting from the difficulty in managing complex systems. The acceleration of scientific development and the inherent difficulties of managing complex systems mean that tragedies, crises, and catastrophes in which technology will be complicit are likely to escalate during the coming decades. When a disaster has occurred, emotions run high and the time for balanced reflection contracts. However, foresight and planning prepares scholars and other interested parties in presenting developed proposals when the time for action is at hand.

My discussion will focus upon societal, ethical, and policy challenges arising from robotics, but will also address a few instances of where robotics and neurotechnologies converge in the form of enhanced humans or techno sapiens. The risks posed by emerging technologies fall within three broad categories:

1. Specific discernible risks that can be largely addressed through innovation, regulation, and soft governance.
2. Far-reaching societal impacts arising from the various ways in which emerging technologies will be combined.

---

[119] Particular attention has been given to synthetic biology, with recent reports from the European Group on Ethics (Capurro et al., 2009) and the Presidential Commission for the Study of Bioethical Issues (Gutmann et al., 2010). The more recent Presidential Commission made a few recommendations but found "…*no reason to endorse additional federal regulations or a moratorium on work in this field at this time*."

3. Existential risks – "*where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential*" (Bostrom, 2002).

These categories are not mutually exclusive, but they are useful for framing the discussion. The existential risks are beginning to be popularised by the media, while most scholars and legal theorists direct their energies toward managing specific risks. Where more attention can and should be directed is upon monitoring and managing the impact of technologies that will be combined in ways that will have far-reaching societal consequences.

Certainty, most of the specific societal challenges arising from the development of robots and neurotechnologies are similar to ethical and legal challenges in other realms for which laws and legislation have already been forged. Manufacturers of present day robots face the same safety and liability concerns that confront companies that market other tools and devices. Testing and distribution of neuropharmaceuticals and neuroprosthetics fall within the canons for research and medical ethics. So in what sense are the ethical and legal challenges posed by robots and neurotechnologies new? This is among the questions I will explore.

I am generally skeptical about contentions that emerging technologies will threaten human existence in the not-too-distant future. But for the purposes of this narrative, I will begin with a discussion of the existential risks, then turn to specific discernible risks, and finish with a discussion of the combinatorial impact of emerging technologies.

Existential Risks

Existential risks are speculative threats such as designer pathogens that could wipe out humanity. The poster children for existential risks associated with robotics are an unfriendly technological singularity and grey goo[120]. Grey goo, first noted by Eric Drexler (1986, chapter 11)) and later popularised by Bill Joy (2000), is a not entirely fanciful notion that self-reproducing nanomachines would eat up all the carbon-based matter leaving the earth covered in a three foot deep sludge of identical tiny machines. The singularity, a time when machines would equal and then exceed human intelligence, was conceptualised by the mathematician I.J. Good (1965), named and developed by the science fiction writer and mathematician Vernor Vinge (1993), and popularised in recent years by the inventor Ray Kurzweil (2006). An unfriendly singularity refers to the concern that computers or robots, which are more intelligent than humans, may not be interested in human welfare. (In the remainder of this article, I will use the spelling '(ro)bots' when referring to both physical robots and intelligent 'bots' within computer networks.) One possibility is that super-intelligent machines will be

[120] Grey goo is a possibility often associated with nanotechnology, as tiny nanomachines or nanobots are likely to be the product of molecular engineering rather than IT.

antagonistic to humans. But a more serious challenge may be actions taken by intelligent (ro)bots in single-minded pursuit of their own goals, which threaten human existence.

Humanity would presumably resist the development of technologies that pose clear existential risks. That of course assumes that (a) we perceive these threats early enough, and (b) we can agree upon an approach for stopping or redirecting the potentially threatening research. There is certainly plenty of hype that AI and other technologies pose near-term (20-100 years) existential risks. But without clear-cut evidence that a particular area of research will cause harms, there will be economic and political pressures to proceed with the development of technologies that offer societal benefits. Speculative risks do not carry much weight in the formulation of public policy. How many of us in either Europe or the U.S., for example, would have been willing to stop all progress in genetics or computer science over the past half century based on 1950's fears of giant locust and robot takeovers?

Furthermore, it will be difficult to forge international agreements for regulating or relinquishing the development of most technologies given the differences in values from country to country. For example, the European Union has codified the precautionary principle[121], while in the American context there tends to be a faith that a 'technological fix' will be available for most, if not all, challenges. Countries with more stringent precautionary policies are likely to be at a competitive disadvantage in reaping the benefits of potentially transformational technologies, while a more open policy could expose the citisenry to the introduction of dangerous products. This, of course, is not a new issue. It informs policy debate in every country as legislatures struggle to balance public safety against economic growth.

There is certainly a need for public education and for the public to engage in a conversation about the longer-term course of technological development. However, it is unclear whether such a conversation can yield practical results within countries with very heterogeneous populations.[122] Competing philosophies, religious beliefs, and cultural narratives will defuse the prospect of formulating clear policy goals. Nevertheless, in a democratic society the public should give at least tacit approval to the futures it is creating.

One serious concern is the likelihood that any discussion of existential risk will be highly politicised. If

[121] Paragraph 2 of article 191 of the Lisbon Treaty (COM, 2000) states that, "Union policy on the environment shall aim at a high level of protection taking into account the diversity of situations in the various regions of the Union. It shall be based on the precautionary principle and on the principles that preventive action should be taken, that environmental damage should as a priority be rectified at source and that the polluter should pay." Since adoption, this communication has come to inform policy beyond the environment including laws related to genetically modified foods and technological development.

[122] Arizona State University (Guston, 2000; Hamlett, Cobb, & Guston, 2009) has experimented with the Danish consensus conference model (Grundahl, 1995) as a tool for representative public engagement spanning the many cultures of the United States.

tacit approval is lacking, tensions will periodically erupt, and could potentially lead to crises that undermine social stability. The debates over the fate of Terri Schiavo and funding for stem cell research were two such mini-crises in the U.S., and may be harbingers of social tensions to come. At their best, such crises are valuable opportunities for public education. When values conflict, the public has an occasion to work through the issues and establish new priorities. At their worst, these crises become politicised in a manner where there is more heat than light. James J. Hughes (2004) argues that the enhancement debate is likely to become a central dividing issue in American politics.

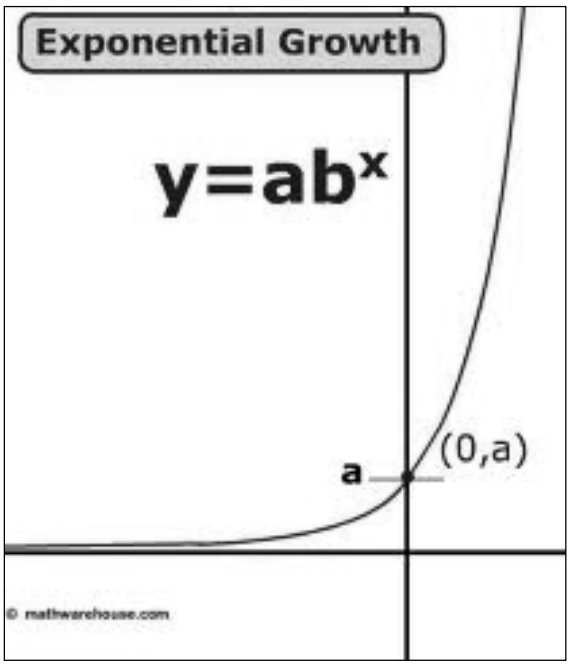

Figure 1: © mathwarehouse.com

In conclusion, reflection upon longer-term existential risks is a fascinating subject that touches upon many of the great philosophical and ethical questions. The conversation can yield broad generally shared societal values and guidelines, but in the absence of a specific threat is unlikely to lead to substantive public policy.

The probable, the plausible and the highly speculative possibilities

The speed at which emerging technologies are developing is a central issue in determining when and if additional mechanisms for their oversight are necessary. There is tremendous confusion regarding which technological possibilities are probable and which are highly speculative. This confusion is rampant not only within the general public, but also among experts. Prognostications as to when, or if, a technological singularity or self-reproducing nanomachines are possible vary from 30 to 200 years or never.

A theory that exponential growth is accelerating the pace of technological development (Kurzweil, 2006) is gaining currency among a vocal community of scientists, futurists, and young adults. The theory focuses on trends where information technologies are expanding exponentially, while costs such as computing capacity and gene sequencing are contracting exponentially. These trends are then projected into the future to make predictions as to what can be expected over the next 5-50 years.

On a line graph, a mild slope represents exponential growth in its early phases over an extended period of time. But eventually the slope shifts upward representing a pattern of accelerating growth.

Certainly not all trends show this accelerated upward shift. But the model of accelerating growth is very difficult to challenge, in that any trend line that does not show a pattern of rapid acceleration may merely represent an early phase of expansion. In other words, exponential change is a theory that is difficult to falsify.

Other theories, such as a somewhat simplistic computational theory of mind, are marshaled to support the view that (ro)bots with human-like cognitive abilities are on the near horizon. By 2013 a supercomputer

will be completed that rivals the computation capacity of the human brain. But there is considerable disagreement as to whether such landmarks signal that robots with human intelligence are in the near-future.

In the more optimistic projections of accelerating change there is a tendency to aggrandise the character and capacities of our presumed evolutionary successors who we will create over the next few decades. This is often accompanied by a tendency to pathologise human nature, a position supported by research in the cognitive sciences that accentuates evolutionarily bequeathed flaws. Together these two positions reinforce an implicit bias that creating our evolutionary successors, whether artificial or biological, is a 'good' in itself.

The academic community tends to be much more skeptical regarding the pace of scientific discovery. To be sure, there are many funding proposals where the prospects for dramatic breakthroughs are hyped. But even in a climate of ongoing scientific progress, the research community remains cognisant of a history of unfulfilled predictions.

Mediating between the more radical visions of futurists who perceive exponential change and the more conservative visions of scholars working in specific fields is incredibly difficult. I share in the skepticism that many of the more dramatic futuristic scenarios will not be possible within the next 20-40 years. But I do believe we need credible mechanisms for monitoring technological developments, and for flagging when thresholds are about to be crossed that hold serious risks for humanity. At present there are no good mechanisms in place to help either the public or scholars discriminate the probable from the plausible, the unpredictable, and the highly unlikely scenarios.

Robotics: Specific ethical and legal concerns

Legal theorists and philosophers have been intrigued for years by the thought experiment of when, or if, future robots might be granted legal rights, or be designated legal persons responsible for their own choices and actions (Lehman-Wilzig, 1981; Solum, 1992; Calverley, 2005). But there is also beginning to be a small body of scholarship that analyses more near-term issues for the robotics industry.

The introduction of robots into the home, the battlefield, and the commerce of daily life poses an array of societal, ethical, legal, and policy challenges. Indeed, limited purpose robots have been proposed for all kinds of human activity. Drones and unmanned ground vehicles developed for the military are being marketed to local police forces. Surveillance drones, some smaller than birds, will be a nightmare for administering the safety of aviation. Driverless cars, cooks, and caregivers are under development. Robots that care for the elderly and homebound are a high priority for countries such as Japan.[123] The array of

[123] Many companies throughout the world are designing service and domestic robots. Robot caregivers will be particularly important to the Japanese where employment is high, the population is aging, and immigration restrictions

applications for robots will also entail a vast array of ethical and legal considerations that must be addressed. The central concerns are subsumed within four interrelated themes: (a) *safety*, (b) *appropriate use*, (c) *capability*, and (d) *responsibility*.

Safety has always been of importance to the engineers who build systems. Existing legal frameworks largely cover the legal challenges posed by present day robots. The robots that have been developed so far are sophisticated machines whose safety is clearly the responsibility of the companies that produce the devices and of the end users who adapt the technology for particular tasks.

Social and ethical theorists have raised a number of questions regarding which tasks are *appropriate* for robots. Some find the use of robots as sex toys offensive. Others lament the sensibilities and lessons lost in substituting robopets and robocompanions for animals or people (Sparrow, 2002; Turkle, Taggart, Kidd, & Daste, 2006). From a humanistic perspectives, turning to robotic caregivers for the homebound and elderly is perceived as abusive or reflecting badly upon modern society, although robotic care is arguably better than no care at all. One dangerous practice is the increasing use of robonannies, robots that tend infants and children. Noel and Amanda Sharkey (2010) argue that the extensive use of robots as nannies, and companions for infants, may actually stunt emotional and intellectual development.

The appropriateness and *ability* of robots to serve as caregivers is commonly misunderstood by the public or misrepresented by those marketing the systems. The limited abilities of present day robotic devices can be obscured by the human tendency to anthropomorphise robots whose looks or behaviour is faintly similar to that of humans. There is a need for a professional association or regulatory commission that evaluates the capabilities of systems and certifies their use for specific activities. This is likely to be very expensive, as the development of each robotic platform is a moving target – existing capabilities are undergoing refinement and new capabilities are constantly being added to systems.

The diminution of privacy and property rights is already a focus for computer ethics and theorists working on information and Internet law. Robots will exacerbate those concerns. For example, introducing robots into the home and other social settings raises privacy risks similar to those posed by surveillance cameras. Robots will have both sensors and large drives that can record all the data they collect. This data offers a benefit in that it can be analysed if anything goes wrong. But it will also be a record of all private activity within range of the sensors. No doubt the hard drives within robots and networks will be subpoenaed for everything from criminal investigations to custody battles. Data stored on robots that are connected to the Internet, as most are likely to be, may be accessible for a variety of criminal purposes (Denning et al., 2009).[124]

---

limit bringing in domestic workers from other countries.

[124] Tamara Denning, Tadayoshi Kohno, Karl Koscher, William Maisel, and colleagues at the University of Washington have already demonstrated that the cameras and sensors in a robotpet can be hacked to gain real-time visual and

Limiting liability

With the increasing complexity of robotic systems, designers and engineers of a device cannot always predict how they will act when confronted with new situations and new inputs. 'Many hands' will have contributed to the building of a robot (Nissenbaum, 1996). The full operation of each hardware component in a system will only be understood by those who designed and built that component, and even they may have little or no understanding of how that component might interact with other components in a totally new device. The pressures to complete projects and the cost of testing also contribute to limited understanding of the potential risks inherent in new devices.

Of course credible manufacturers do not want to be held liable for marketing faulty devices. They may elect to avoid releasing products whose safe use they have no way of guaranteeing. For a society banking on the productivity improvements that transformative technology such as robots offer, this could be perceived as a heavy burden on innovation and a heavier price to pay for systems whose risks are low but whose benefits are significant. Indeed, other countries with higher bars to litigation would be likely to take a lead in robot technologies as manufacturers wait for liability law to be sorted out in their own country.

Manufacturers will certainly welcome measures that lower their liability. As a means of spurring industry growth and innovation, Ryan Calo (2010) has proposed immunising manufacturers of open robotic platforms from all actions related to improvements made by third parties. But any approach to limiting liability must be balanced against insuring that industry does not knowingly introduce dangerous products.

No-fault insurance for robots is another approach that could lower manufacturer's liability. Consider driverless cars, such as the one Google has developed. Even if driverless cars are much safer than those driven by humans, robot-chasing attorneys are likely to initiate suits for each death in which a robotic car is involved. All new technologies face similar challenges. Free societies have an array of laws, regulations, insurance policies, and precedents that help protect industries from frivolous lawsuits. Companies pursuing the huge commercial market in robotics will protect their commercial interests by relying on the existing frameworks and by petitioning legislatures for additional laws that help manage their liability.

Difficulty in establishing responsibility for harms

The Challenger disaster is a case study in the difficulty of determining cause and responsibility for the failure of complex systems. Millions of dollars were spent before investigators established that the culprit was the brittleness of tiny o-rings in cold weather. Later, investigators questioned whether precautionary

---

auditory access to activity in the pet's location. More alarmingly, they have hacked and altered heart pacemakers (Maisel & Kohno, 2010) and the software in automobile computers that regulates braking and other functions (Koscher et al., 2010). Kohno believes that every topic in computer science can have a security-related twist.

measures would have uncovered a flaw like this in the design of the system. In reviewing that research, Malcolm Gladwell writes, "*we have constructed a world in which the potential for high-tech catastrophe is embedded in the fabric of day-to-day life*" (1996).

Manufacturers will encourage an appreciation for the difficulty in establishing responsibility for complex intelligent systems as a way of diluting or mitigating liability for system failures. Simultaneously, practical ethicists and social theorist are raising concerns as to the dangers inherent in diluting corporate and human responsibility, accountability, and liability for the actions of increasingly autonomous systems. Recently, five rules have been proposed as a means of reestablishing the principle that humans cannot be excused from moral responsibility for the design, development, or deployment of computing artefacts.[125]

Rule 1: The people who design, develop, or deploy a computing artefact are morally responsible for that artefact, and for the foreseeable effects of that artefact. This responsibility is shared with other people who design, develop, deploy or knowingly use the artefact as part of a sociotechnical system.

Rule 2: The shared responsibility of computing artefacts is not a zero-sum game. The responsibility of an individual is not reduced simply because more people become involved in designing, developing, deploying or using the artefact. Instead, a person's responsibility includes being answerable for the behaviours of the artefact and for the artefact's effects after deployment, to the degree to which these effects are reasonably foreseeable by that person.

Rule 3: People who knowingly use a particular computing artefact are morally responsible for that use.

Rule 4: People who knowingly design, develop, deploy, or use a computing artefact can do so responsibly only when they make a reasonable effort to take into account the sociotechnical systems in which the artefact is embedded.

Rule 5: People who design, develop, deploy, promote, or evaluate a computing arteact should not explicitly or implicitly deceive users about the artefact or its foreseeable effects, or about the sociotechnical systems in which the artefact is embedded.

The rules are broad in scope. Their application would not be easy, and might significantly slow the development of the robotics industry. For example, the rules propose that those who develop and market computing artefacts are morally responsible for the foreseeable ways in which the artefact might be used. If this standard were codified in law, the manufacturer of a gun would be accountable for its use in a murder, and the manufacturer of cigarettes accountable for lung cancer.

Whether rules, such as those proposed for computing artefacts can or should be translated into liability law remains an open question. There is a difficult policy debate ahead. Should accountability and liability for

[125] The full document titled, *Moral Responsibility for Computing Artifacts*, defines terms and explains the rules. It can be accessed at https://edocs.uis.edu/kmill2/www/TheRules/.

computing artefacts be lowered in order to stimulate the development of a potentially transformational industry? Or, should existing protections be maintained even if this arrests the willingness of companies to introduce products that offer significant benefits with low or uncertain risks?

Placing decisions made by (ro)bots ahead of human intelligence is a mistake. While IT systems may exceed human intelligence in some aspects such as searching large databases, they are a long way off from matching human intelligence in so many other dimensions. Unfortunately, humans are uncomfortable in going against the recommendations of semi-intelligent systems. Human decision makers need to be empowered when they have the courage to override the actions or suggestions of robotic systems. Claims that robots have the capabilities to make superior decisions, or even function as safe substitutes for human agents in social contexts should be examined skeptically.

Moral machines

If robots can be designed so that they are sensitive to ethical considerations and factor those considerations into their choices and actions, new markets for their adoption will be opened up. However, if they fail to adequately accommodate human laws and values, there will be demands for regulations that limit their use.

A new field of inquiry variously known as machine ethics, machine morality, computational ethics, artificial morality, and friendly AI has emerged in response to the advent of increasingly autonomous (ro)bots. When designers and engineers can no longer anticipate how intelligence systems will act when confronted with new situations and new inputs, it becomes necessary for the (ro)bots themselves to evaluate the appropriateness or legality of various courses of action.

Machine ethics (ME) should be distinguished from robot ethics. While the latter addresses societal concerns in the deployment of robots, ME considers the prospects for developing machines that are explicit moral reasoners. Initially, (ro)bots capable of making moral decisions will function within contexts where their freedom of action is limited. However, as autonomy increases (ro)bots may eventually evolve into artificial moral agents (AMAs hereafter). But there are many issues as to whether (ro)bots can acquire the full intelligence and moral acumen to actually be considered true moral agents. Many thresholds, both technological and philosophical, must be crossed between here and there. Some of the thresholds looming may turn out to be ceilings that define limits to the intelligence and moral understanding of (ro)bots.

Operational morality and appropriate behaviour

The chart below will be helpful for appreciating the development of (ro)bots as autonomy and sensitivity to moral considerations expands.

All technology can be viewed as falling within this chart. A hammer has neither sensitivity nor autonomy. A thermostatic has some sensitivity to temperature and the autonomy to turn a furnace or fan on or off when a threshold has been reached.

Most of the robotic devices available or being developed are operationally moral in the sense that the corporations and engineers who build the device determine the values instantiated in the robots actions and

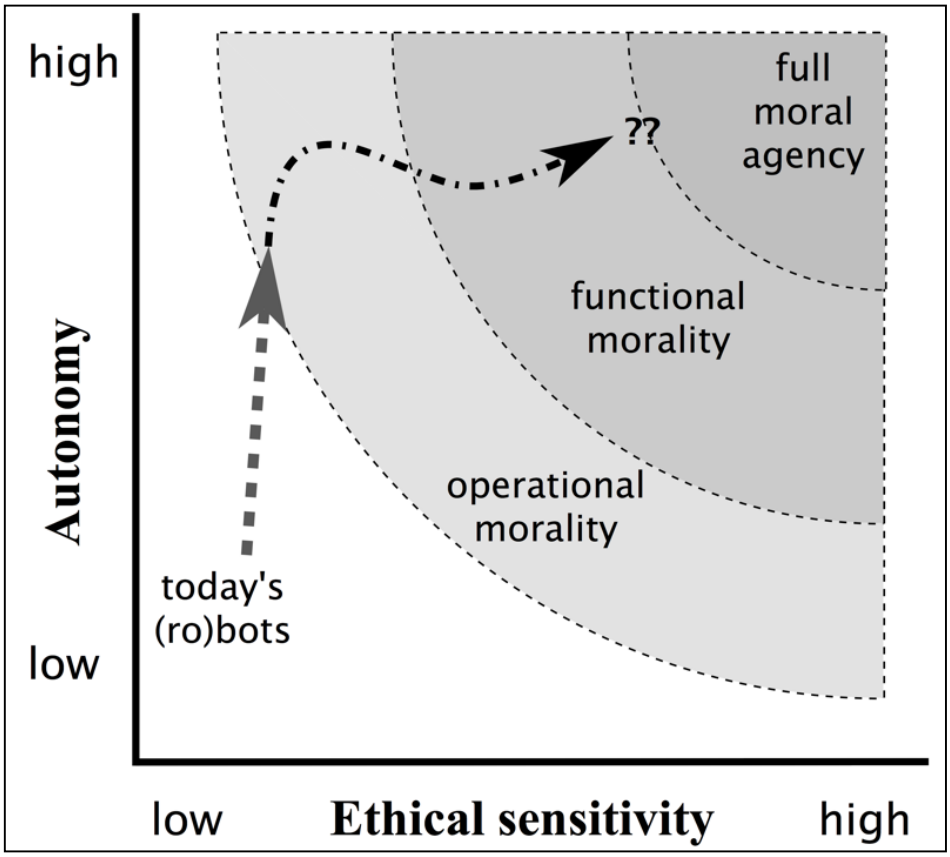

Figure 2: Plotting ethical sensitivity and autonomy in robots

choices. Proscribed behaviour is programmed into the systems.

Appropriate behaviour

What kinds of behaviour are appropriate for a robot? Whose values or what values should be instantiated in a robotic device? A few years back the manufacturer of a speaking robot doll considered what the doll should do if treated abusively by a child. The engineers knew how to build sensors into the doll that would alert the system to such behaviour. After analyzing the issues and consulting with lawyers they decided that the doll would say and do nothing.

How should a robot caregiver perform in an ethically charged situ-ation? What should a robot do if an elderly patient refuses medi-cine?[126] Or, what if the robot enters a room and dis-covers the person under its supervision is hysterical? How would the robot know that the fear on the face of its charge was caused by the robot itself or by some other factor?

Consider a robot that is the companion of a young child or teenager. Should the robot intervene if the child puts itself in harms way? Are there circumstances where inappropriate intervention by the robot might do some harm? Would programming a robot to tell a child to stop abusing himself be a good or a bad idea? What if the child ignores the directive? Should the robot discipline the child? A robot that instructs but cannot follow up with discipline may well be teaching the wrong lessons. But a robot that disciplines is not likely to

[126] Michael and Susan Anderson (2008) have addressed the ethical considerations regarding what to do if a patient refuses medicines in designing software that they have implemented in both a computer system and in a robot.

instill trust.[127]

How should a robot nanny respond to a child that relates to the nanny in a way that would be inappropriate or even physically violent if the nanny were human? Would you want the robot to say, 'Stop that! You are hurting me,' presuming (as is probable) that the robot has no capacity to feel pain? While well intentioned, such a statement by a robot is absurd, and could lead to unintended consequences. There are countless similar situation that could arise.

Robots will be able to mechanically discern certain ethical challenges, presuming that the designers and engineers anticipate the challenge and program in an appropriate response. But one response may not suit everyone. Some parents might want a robot to tell a youngster to 'stop' if she is relating to the robot in ways that would hurt a human. Other parents would reject having a robot dispense a reprimand. Software could be designed and implemented that made it a user option (parental choice) as to the manner in which a robot caregiver would respond to a child in ethically charged situations. During setup, parents would be introduced to a variety of ethically charged situations. They could be informed about the ramifications of different alternatives, and the responsibility they were taking on in placing the child in situations where a robot might need to take such actions. This proposal requires further thought, but a setup procedure would provide an excellent opportunity for manufacturers of companion robots to educate parents on what they could and could not expect from such devices. The parents get a little education on the proper use of robonanny, and the manufacturer protects itself from certain forms of liability.

The plethora of such new ethical challenges will hopefully alert leaders of industry to the importance of making ethical considerations an integral aspect of the design process. It is heartening that schools of engineering have gone beyond giving lip service to professional ethics, and have become truly concerned with ensuring that their students are sensitised to the societal impact of the products they design. The next step lies in applied ethicists joining the design process, not as naysayers, but as members of the team looking for ways to engineer solutions to societal and ethical challenges. Philosopher Helen Nissenbaum (2001) calls this 'engineering activism'.

Functional morality

Robots capable of even limited autonomous activity will need to factor an array of considerations in determining what behaviour is appropriate or legal when confronted with difficult ethical challenges. The field of machine morality is largely concerned with the approaches and procedures used by the (ro)bot to make

[127] A version of this discussion and the remaining paragraphs in this section appeared in an article (Wallach, 2010) published in the *Journal Interaction Studies*. The essay was a response to a target article by Noel and Amanda Sharkey (2010).

such judgments. I have written about this subject extensively, so I will only mention a few brief details here.[128]

The approaches for implementing moral decision-making capabilities in robots fall within two broad categories, top-down and bottom-up (Allen, Smit, & Wallach, 2006). Top-down refers to the implementation of rules, principles, or moral decision-making procedures, such as utilitarianism, Kant's categorical imperative, the Ten Commandments, Hinduism's yama and niyama, and even Asimov's laws. A top-down approach takes an antecedently specified ethical theory and analyses its computation requirements to guide the design of algorithms and subsystems capable of implementing the theory. Bottom-up approaches take their inspiration from evolutionary psychology and game theory, as well as developmental psychology and theories of moral development. Bottom-up approaches, if they use a prior theory at all, do so only as a way of specifying the task for the system, but not as a way of specifying an implementation method or control structure.

Both top-down and bottom-up approaches have their strengths and weaknesses. For example, principles defined broadly can cover countless situations, but if too broad or too abstract their application to specific challenges will be debatable. Bottom-up approaches are particularly good at dynamically integrating input from discrete subsystems. But defining the ethical goal for a bottom-up system would be difficult, as would assembling a large number of discrete components into a functional whole.

Eventually, we may need AMAs that maintain the dynamic and flexible morality of bottom-up systems that accommodate diverse inputs, while subjecting the evaluation of choices and actions to top-down principles that represent ideals we strive to meet. Furthermore, AMAs will require additional capabilities in order to be sensitive to a range of ethical considerations or to acquire access to essential information. These supra-rational capabilities (beyond reason) include emotions, social intelligence, a theory of mind, empathy, consciousness, and being embodied in a world with humans, objects, and other agents.

There are many questions as to whether all these capabilities can be instantiated computationally. But as the sensitivities and abilities of robots expand, new applications for the use of robots will open up. One of the tasks for machine ethics is to delineate the capabilities AMAs will require in order to operate appropriately and safely within specific domains.

The task of building AI systems with moral decision-making capabilities can be understood as encompassing two hard problems. The first problem entails finding a computational method to implement norms, rules, principles, or procedures for making moral judgments. The second is a group of related challenges that I refer to as frame problems. How does the system recognise it is in an ethically significant

[128] See Wallach, W. & Allen, C. (2009). *Moral Machines: Teaching Robots Right From Wrong*, New York: Oxford University Press, for the most comprehensive overview of the topic.

situation? How does it discern essential from inessential information? How does the AMA estimate the sufficiency of initial information? What capabilities would an AMA require to make a valid judgment about a complex situation, e.g., combatants v. non-combatants? How would the system recognise that it had applied all necessary considerations to the challenge at hand or completed its determination of the appropriate action to take? For example, what stopping procedure would the system use to determine that it had completed a utilitarian calculation?

The development of AMAs is likely to be a long, incremental process. Throughout this development, a primary challenge for society will be the monitoring and assessing of the capabilities of each system. What criteria should be used to determine whether a particular system could be deployed safely in a specific context? What oversight mechanisms need to be put into place in order to ensure that such an assessment can be made and has been made? What penalties might be applied if a certified system is later implicated in harmful actions?

If sophisticated AMAs can be built, the more distant theoretical and speculative challenges that have fascinated science fiction writers, philosophers, and legal theorists will come into play. Will artificial agents need to emulate the full array of human faculties to function as adequate moral agents and to instill trust in their actions? What criteria should be used for evaluating whether an AI system deserves rights or should be held responsible for its own actions? Does punishing a robot make any sense? If yes, how might one punish a robot for infractions of rules or transgressions against the rights of others? Should or can we control the ability of robots to reproduce? How will humanity protect itself against treats by intelligent (ro)bots?

Combinatorial risks and societal impact

If only a fraction of the technologies being proposed come to pass within the next decades, human behaviour and human culture will be transformed dramatically. By our standards people alive at the end of the 18th century were superstitious, unscientific, unsanitary, provincial, and filled with racial, sexual, and class prejudices. But humanity was about to be transformed by the industrial revolution, a germ revolution in medical science, and by the sanitation revolution. The changes in the next thirty years may be as dramatic as the changes over the past two hundred years.

Three related examples will suffice to illustrate highly probable trends that will have profound social impacts. These examples fall far short of the existential risks stirred up by speculative possibilities. Most of the technologies I will mention have already been developed, although a few may not be practical. They do, however, illustrate societal and ethical challenges that may well arise, but will not necessarily be addressed by the kinds of policy mechanisms that are presently in place.

Technological unemployment

> In quite a few years – in our own lifetimes I mean – we may be able to perform all the operations of agriculture, mining, and manufacture with a quarter of the human effort to which we have been accustomed… We are being afflicted with a new disease of which some readers may not yet have heard the name, but of which they will hear a great deal in the years to come – namely, technological

> unemployment. This means unemployment due to our discovery of means of economising the use of labour outrunning the pace at which we can find new uses for labour.
>
> But this is only a temporary phase of maladjustment. All this means in the long run that mankind is solving its economic problem. I would predict that the standard of life in progressive countries one hundred years hence will be between four and eight times as high as it is today. There would be nothing surprising in this even in the light of our present knowledge. It would not be foolish to contemplate the possibility of a far greater progress still. (Keynes, 1930)

Contentions that technology and industrialisation would undermine employment opportunities predate Keynes' prediction. But in each generation human ingenuity has repeatedly generated new forms of employment. Yet the specter of technological unemployment is revisiting us once again with what appears to be the second jobless recovery in a decade. Ingenuity and the desire to work may again create employment opportunities. But two factors alter the present equation: (a) ever-increasing life expectancy and later retirement, and (b) the development of robots capable of performing an ever-increasing number of intelligent tasks.

Average life expectancy is growing at a rate of 2-4 years each decade in advanced industrial countries. Radical life extension is among the possibilities that are being discussed. Aubrey de Grey has suggested that it may even be possible to end death (de Grey & Rae, 2007). To date, however, there appears to be a ceiling on overall life expectancy at around 120 years of age. But many new subfields, including stem-cell research, personalised medicine, and RNA interference, will contribute to regenerative medicine. Even if the rate of growth for average life expectancy remains at its present pace, delays in retirement and benefits to those who do retire and live longer will lead to serious tensions within the social fabric.

Pressures on the job market are also exacerbated by intelligent machines capable of performing an increasing number of tasks at lower cost than human labor. If the human workforce actually contracts as the population expands (births plus extended lives), society will need to develop new mechanisms for the distribution of capital, goods, and services.

Certainly biotechnologies that extend life will be popular and widely adopted. But should governmental policy be directed at extending life? Or, should extending life be treated as secondary to other goals, such as ensuring the quality of life for all citisens up to a particular age, e.g., 82 years?

Cyborg warriors

The military is a driving force in speeding up the development of many new technologies. The goal of military planners is to achieve a strategic advantage in warfare. Little attention is given to the broader societal impact of technologies financed for military use. The scientists on the Manhattan Project did not understand how atomic weapons would radically alter a world under constant threat of their use. There is little or no reflection on how the strategic advantage achieved by roboticising aspects of warfare is likely to be far outweighed by its long-term consequences.

One active trajectory in military research is combining technologies to create the enhanced soldier or

cyborg warrior. Future soldiers are likely to be outfitted with robotic exoskeletons that enhance strength and stamina. Neuroprosthetics devices that convey thoughts might facilitate their communicating with other members of the team. Insect sized drones fly around the battlefield looking for guerilla fighters and beam back their position, which are then overlaid on the soldier's visor. A tablet computer built into a data glove can be used to direct larger weapons-carrying drones. Nanosensors on the body and in the bloodstream will facilitate supervisors, not engaged in combat, in monitoring the physiological well being of soldiers during a skirmish.

Before going to the frontline a soldier could be given a cocktail of modafinil to improve attention, the latest cognitive enhancers that heighten memory and speed up reaction time, and propranolol to reduce the possibility of post-traumatic stress disorder (PTSD).[129] The likelihood of soldiers exposed to the stresses of high-speed combat developing PTSD will be of particular concern to military planners. A combatant will have been pre-screened for biomarkers that indicate great resilience if he should experience a traumatic event. For example, a study conducted by Elisabeth Binder and colleagues found an association between polymorphisms of the FKBP5 gene, childhood abuse, and the risk of an adult getting PTSD (Binder et al., 2008). How many more risk factors for PTSD will need to be revealed before scientists can screen out with a high degree of probably those individuals most susceptible to PTSD if exposed to constant stressors? What probability would justify shielding that individual from stressful occupations or environments: 50%? 85%? Should only those with high resilience profiles be allowed on the frontline during warfare? Will those who have lower resilience profiles be barred from the police force or from fighting fires?

Many societies, from ancient Sparta to the U.S. Marine Corps, have cultivated their warrior class. However, explicitly limiting combat to individuals with a very specific profile has profound ramifications for a democratic society. Furthermore, applying screening techniques for social engineering purposes can lead to new forms of discrimination. On the other hand, given the long-term suffering and the costs to society for the healthcare of veterans with combat-related PTSD, is it irresponsible to send those with a higher risk profile into the theater of war?

Techno sapiens

Students and early adopters are already engaged in widespread experimentation with supplements and prescription drugs in hopes of getting a competitive advantage or for recreational purposes. As new cognitive enhancing drugs become available, one can presume that they will also be combined with a wide variety of other pharmaceuticals. What will be the responsibility of governmental agencies, insurance

[129] While it has been hypothesised that taking propranolol prophylactically might minimise guilt or susceptibility to PTSD, the theory has not been proven.

companies, and educational institutions for adverse incidents arising from using cognitive enhancers for purposes for which they were not specifically prescribed? Will government and non-governmental agencies have the resources to track which of these combinations cause side effects, mental distress, or neurological damage?

Pharmaceuticals are already commonly prescribed in combinations that have never been tested. Arguably we already live in a pharmaceutical regime where multiple drugs are taken to compensate for side effects created by other drugs that have been prescribed. If the harmful consequences of widely disseminated cognitive cocktails do not show up in the short-term, there will be future crises to manage.

Cognitive enhancers will also be combined with both exogenous and endogenous devices, including glasses that augment reality and neuroprosthetics that have been developed for both military and therapeutic purposes. Tools for surfing the web and interfacing with computer devices through thoughts or small muscular movements will be particularly popular. These computer interfaces might be used to manage devices at a distance, on the body, or in the bloodstream.

We also have no understanding whether combining various cognitive enhancers with neuroprosthetics will optimise the freedom of individuals or undermine their autonomy. The mind is a delicate instrument. Optimising one capability could easily interfere with another. Just as texting while driving is a dangerous enterprise, so too may mixing various drugs and tools that enhance individual skills. There should be no tolerance for technologies that undermine the capacity of individuals to be responsible for their actions.

An opportunity or a threat to humanity?

Values differ, as do perceptions of whether the transformation of humanity by emerging technologies is good or bad. If large segments of the public find alterations in human nature, character, or presentation (e.g., cyborgs) offensive, all technologies that transform human identity might be evaluated as being existential threats. There is the rather melodramatic possibility that we are inventing the human species as we have known it out of existence.

Nevertheless, it is hard to imagine anything, short of a disaster that empowers Luddite political parties, arresting technological development. After all, most of the enhancements that cause disquiet within some communities will result as byproducts of research that serves therapeutic needs. A political platform which declares that the lame will never walk or those suffering from trauma will never live normal lives is unlikely to receive widespread acceptance.

Research on intelligent robots and enhancements will also be furthered in the name of economic productivity and personal or corporate freedom. The outstanding question is whether some limits can or should be placed on the development of robotics and other emerging technology? Agreed upon limits might quell concerns that technological development is out of control or becoming the primary force shaping humanity's destiny. But more importantly, some limits will help stave off technological errors that cause harms to the public in the form of economic disruption, environmental degradation, a major health disaster, or political turmoil.

Monitoring and managing emerging technologies

The likelihood that new technologies will be combined in ways that are difficult to predict poses some very tricky policy challenges. How can each society, and humanity as a whole, monitor and manage technological development when the tools we have for forecasting and risk assessment are highly subjective? Certainly any roadmap would be a work in progress as the possibilities change with each new scientific discovery.

As a first project, I would like to propose the creation of a credible vehicle for monitoring and evaluating the state of technological development. Making such a proposal is much easier than knowing how such a vehicle should be designed or implemented.

Government institutions tend to be shortsighted. The EU has taken perhaps more initiative than other governments in convening advisory committees and financing research directed at formulating policy for managing emerging technologies. Foundations and other funding sources have been slow to perceive this subject as one where their grants will nurture effective research. Universities give lip service to fostering interdisciplinary research, but few actually reward scholars for interdisciplinary work. Within academia the prevailing presumption is that more general or comprehensive research lacks rigour or empirical foundations. Nevertheless, there is a need, and with some effort and attention that need will come to the fore. Hopefully it will be possible to generate that attention without a serious crisis in which an emerging technology has been complicit.

A first stage in the development of a credible vehicle for monitoring emerging technologies might be a series of expert workshops. Expert conferences are not just opportunities to present ideas and educate each other, but should also be designed to provide an occasion to grapple with specific issues and debate possible solutions. The challenges are largely beyond the ability of one individual to grasp, but there are many scholars, engineers, leaders of industry, and policy planners who have expertise on essential aspects of those challenges. For example, William Halal's TechCast project has been periodically sampling one hundred experts on when projected technologies will appear (Halal, 2008). Leon Fuerth, formerly Al Gore's National Security Advisor, has proposed how the executive branch of the U.S. government can be reorganised to accommodate anticipatory planning, what Fuerth calls 'forward engagement' (2009). Futurists like Dennis Bushnell, chief scientist for NASA Langley, have begun trying to think through the potential combinatorial impact of current technological, economic, and environmental trends. There is also no shortage of bioethicists with interdisciplinary expertise on challenges posed by the tech revolution.

A few of these expert workshops would brainstorm models for think tanks, research centers, or governmental agencies whose reports would be considered credible and worthy of attention by both the general public and other experts. Other expert workshops would focus on related issues. Among the topics that should be given more attention:

1. *Existing Policy Mechanisms*: Are they adequate for managing the combinatorial impact of emerging technologies?
2. *Public Education*: How do you get an informed public? Would the Danish model for informed citisen output early in the process of developing emerging technologies work in larger countries

such as the U.S.?

3. *Monitoring the Speed of Technological Development*: Can we defuse some of the anxiety around emerging technologies by monitoring which technological thresholds are likely to be crossed within the next five to ten years?
4. *Kinds of Control*: Slowing or thwarting scientific research is difficult in any circumstance. What kinds of leverage are there within existing policy mechanisms for modulating the societal impact of emerging technologies? When will weakening a particular mechanism undermine the entire safety net of public protections? Will the introduction of additional policy mechanisms significantly alter the ability to manage emerging technologies?
5. *Managing Complexity*: Are there problems in managing complex systems that make periodic crises inevitable?
6. What kind of planning for '*Black Swans*,' low probably high impact events makes senses (Taleb, 2007).
7. *Crisis management*: What kinds of crises should we plan for, and would such preparation be money well spent?
8. What new strategies for *comprehensive risk assessment* can be developed?
9. *International considerations*: Will international considerations thwart national attempts to modulate the harms posed by emerging technologies? Which concerns are shared?
10. *The Downside of a Comprehensive Approach*: Differing projections of what is probable are likely to be politicised by both those who support and those who wish to halt the implementation of new technologies. How might efforts to politicise discussion be defused?
11. What are the *primary values* that should inform technology policy?

Conclusions

We are collectively in a dialogue directed at forging a new understanding of what it means to be human. Pressures are building to embrace, reject, or regulate robots and technologies that alter the mind/body. How will we individually and collectively navigate the opportunities and perils new technologies offer? With so many different value systems competing in the marketplace of ideas, what values should inform public policy? Which tasks is it appropriate to turn over to robots and when do humans bring qualities to tasks that no robot in the foreseeable future can emulate? When is tinkering with the human mind or body inappropriate, destructive, or immoral? Is there a bottom line? Is there something essential about being human that is sacred, that we must preserve? These are not easy questions.

Among the principles we should be careful not to compromise is that of the responsibility of the individual human agent. In the development of robots and complex technologies those who design, market, and deploy systems should not be excused from responsibility for the actions of those systems. Technologies that rob individuals of their freedom of will must be rejected. This goes for both robots and neurotechnologies.

Just as economies can stagnate or overheat, so also can technological development. The central role

for ethics, law, and public policy in the development of robots and neurotechnologies will be in modulating their rate of development and deployment. Compromising safety, appropriate use, and responsibility is a ready formulation for inviting crises in which technology is complicit. The harms caused by disasters and the reaction to those harms can stultify technological progress in irrational ways.

It is unclear whether existing policy mechanisms provide adequate tools for managing the cumulative impact of converging technologies. Presuming that scientific discovery continues at its present relatively robust pace there may be plenty of opportunities yet to consider new mechanisms for directing specific research trajectories. However, if the pace of technological development is truly accelerating the need for foresight and planning becomes much more pressing.

References

Allen, C., Smit, I., & Wallach, W. (2006). Artificial Morality: Top-Down, Bottom-Up and Hybrid Approaches. *Ethics and New Information Technology, 7*, 149-155.

Anderson, M. & Anderson, S. (2008). EthEL: Toward a principled ethical eldercare robot. *ACM/IEEE Human-Robot Interaction Conference*, Amsterdam.

Binder, E.B., Bradley, R.G., *et al.* (2008). Association of FKBP5 polymorphisms and childhood abuse with risk of posttraumatic stress disorder symptoms in adults. *JAMA, 299*(11), 1291-1305.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology, 9*(1).

Calverley, D. (2005). Towards a method for determining the legal status of a conscious machine. Paper presented at the *Artificial Intelligence and the Simulation of Behavior '05: Social Intelligence and Interaction in Animals, Robots and Agents Symposium on Next Generation Approaches to Machine Consciousness*. Hatfield, UK.

Capurro, R., Kinderlerer, J., da Silva, P.M., & Rosell, P.P. (2009). Opinion no. 25: Ethics of synthetic biology, European Group on Ethics. Retrieved from http://ec.europa.eu/european_group_ethics/docs/opinion25_en.pdf

de Grey, A. & Rae, M. (2007). *Ending aging: The rejuvenation breakthroughs that could reverse human aging in our lifetime.* New York:St. Martin's Press.

Denning, T., Matuszek, C., Koscher, K., Smith, J.R., & Kohno, T. (2009). A spotlight on security and privacy risks with future household robots: Attacks and lessons. *Proceedings from the 11th International Conference on Ubiquitous Computing*, Orlando, Florida.

Drexler, E. (1986). *Engines of Creation.* Bantam Doubleday Dell.

Fuerth, L. (2009). Operationalizing forward engagement: Toward anticipatory governance. Retrieved from http://www.forwardengagement.org/storage/forwardengagement/documents/fuerth-_operationalizing_forward_engagment_-_toward_anticipatory_governance_1.20.10final.pdf

Gladwell, M. (1996). Blowup. *The New Yorker*. Reprinted in, *What the dog saw: And other adventures.* Little, Brown and Company, 2009.

Good, I.J. (1965). Speculations concerning the first ultraintelligent machine. *Advances in Computers*, *6*.

Grundahl, J. (1995). The Danish consensus conference model. In S. Joss, & J. Durrant (Eds.), *Public participation in science: The role of consensus conferences in Europe*. Science Museum.

Guston, D. (2010). The anticipatory governance of emerging technologies. *Journal of Korean Vacuum Society, 19*(6), 432-441.

Gutmann, A., Wagner, J., *et al* (2010). New directions: The ethic of synthetic biology and emerging technologies. Presidential Commission for the Study of Bioethical Issues. Retrieved from http://www.bioethics.gov/documents/synthetic-biology/PCSBI-Synthetic-Biology-Report-12.16.10.pdf

Hughes, J. (2004). *Citizen cyborg: Why democratic societies must respond to the redesigned human of the future.* Westview Press.

Halal, W. E. (2008). *Technology's promise: Expert knowledge on the transformation of business and society.* Palgrave Macmillan.

Hamlett, P., Cobb, M.D., & Guston, D. (2009). National citizens' technology forum report. Retrieved from http://cns.asu.edu/files/NCTFSummaryReportFinalFormat08.pdf

Joy, B. (2000). Why the future doesn't need us. *Wired, 8*.

Keynes, J.M. (1963 [1930]). Economic possibilities for our grandchildren. In, *Essays in Persuasion* (358-373). W.W. Norton & Co.

Koscher, K., Czeskis, A., Roesner, F., Patel, S., Kohno, T., Checkoway, S., *et al.* (2010). Experimental security analysis of a modern automobile. *IEEE Symposium on Security and Privacy*, Berkeley, CA, May 16-19.

Kurzweil, R. (2006). *The Singularity is Near*. New York: Viking Penguin.

Lehman-Wilzig, S. (1981). Frankenstein Unbound: Towards a legal definition of Artificial Intelligence. *FUTURES*, 442-457.

Maisel, W.H., & Kohno, T. (2010). Improving security and privacy of implantable medical devices. *New*

England Journal of Medicine, 362(13), 1164-1166.

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics, 2*, 25-42.

Nissenbaum, H. (2001). How computer systems embody values. *Computer, 3*(3), 117-120.

Polyani, M. (1951). *The Logic of Liberty*. University of Chicago Press.

Popper, K. R. (2002 [1945]). *The Open Society and Its Enemies: Vol 1 and 2*. Routledge.

Sharkey, N. & Sharkey, A. (2010). The crying shame of robot nannies: An ethical appraisal. *Interaction Studies, 1*(2), 161-190.

Solum, L.B. (1992). Legal personhood for artificial intelligences. *North Carolina Law Review, 70*, 1231.

Sparrow, R. (2002). The march of the robot dogs. *Ethics and Information Technology*, *4*(4), 305-318.

Taleb, N.N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

Turkle, S., Taggart, W. Kidd, C.D., & Daste, O. (2006). Relational artifacts with children and elders: The complexities of cybercompanionship. *Connection Science, 18*(4), 347-361.

Wallach. W. (2010). Applied ethicists: Naysayers or problem solvers? *Interaction Studies*, *11*(2), 283-289.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from Wrong*. New York: Oxford University Press.

____ Communication from the Commission on the precautionary principle (2000). Consolidated version of the treaty on the function of the European Union.

Chapter 12
Taking the moral stance: Morality, robots, and the intentional stance

Samir Chopra
Brooklyn College and Graduate Center of the City University of New York
Department of Philosophy
✉ schopra@sci.brooklyn.cuny.edu

Abstract The theory of the intentional stance often serves as the basis of arguments for the coherence of ascribing intentional attitudes to a variety of non-human entities. For instance, it may be said that corporations are intentional agents because it is possible to reliably, voluminously and successfully, predict their behaviour using a set of intentional generalisations. Thus, if we analyse the possession of a moral sense to be contingent on the possession of, and rational acting upon, a privileged set of beliefs and desires i.e., the moral ones, we have a means for ascribing a moral sense to an artificial agent such as a robot. For note that our interpretation of human beings as moral agents is dependent on our adopting a 'moral stance' towards them: we ascribe a moral belief ('John believes helping the physically incapacitated is a good thing') and on the basis of this ascription, predict actions ('John would never refuse an old lady help') or explain actions ('He helped her cross the street because he wanted to help a physically incapacitated person'). To display a moral sense is to provide evidence of the direction of action by a set of beliefs and desires termed moral. If we could predict an artificial agent's behaviour on the basis that it rationally acts upon its 'moral' beliefs and desires, the adoption of such a moral stance towards it is a logical next step. Ostensible failures of morality on the part of artificial agents could be understood as failures of reasoning: the failure to hold certain beliefs or desires, or to act consistently with those beliefs and desires. The use of a language of morally-inflected beliefs and desires in successfully, reliably, and voluminously predicting the behaviour of an artificial agent renders plausible a description of an artificial agent, such as a robot, as a moral agent.

Keywords robots, moral agents, intentional stance, moral stance, intentional agents.

As usual the third-person point of view makes progress while the first-person point of view peters out into a systematically mysterious question about imagined intrinsic properties. (Dennett, 1987, p. 107)

Introduction

Can robots be moral? I want to argue this question is most perspicuously framed as the question of whether a robot can be a moral *agent*.[130] Further, as the ascription of agency is dependent on the successful ascription of an appropriate set of beliefs and desires to a putative *intentional entity*, and because a moral agent is a kind of intentional agent, robots can be considered moral agents if they are reckoned as intentional agents displaying direction of their actions by a set of beliefs and desires termed moral. The best strategy for such ascriptions is that of the intentional stance.

Such an ascription of morality is done, much like, though not exactly, the way HLA Hart suggested we ascribe a legal system to a societal grouping, via the *recognition* of a set of operative rules in a group whose violations are responded to with approbation and disapproval. This is a view that is deflationary as far as the metaphysical pretensions of morality are concerned, but it does not diminish any of its normative weight.

Thus, my claim is that the ascription of morality to robots depends (just like it does in the case of human beings) on the *identification and ascription of moral mental states*. To accomplish this we should draw on the well-established battery of techniques of folk psychology, ignore worries about internal constitution, subjective perspectives, and intrinsic properties, and concentrate on linguistic assertions and behavioural evidence. Quine and Davidson made famous the field linguist's task of constructing translation manuals to determine a foreign race's beliefs and language. I draw upon the image of a *field moralist* studying aliens to determine whether they have morality akin to ours. The field moralist's best resource will be the framework of agency, intentionality, and rationality provided by the intentional stance. As such, she should engage in *moral folk psychology*, the most perspicuous strategy available to her. And to us, as we prepare for our encounter with the morality of robots.

Agency and the Intentional Stance

'Agent' has an intuitive meaning in everyday life: something able to take actions. Agents *do* things, they *act*, as opposed to have things *happen to them*; the actions of the agent distinguish it from the rest of the world; to deny something agency is to deny it the capacity to take actions.

This characteristic of agents is indispensable for the identification of actions in our world; to identify an action is to identify an agent as its cause. Or, rather, 'intentional agent,' for actions are not genuine actions

[130] I do not consider whether robots can be persons; however, the challenge of ascribing intentionality is similar. For personhood, a necessary condition might be that "*X has, or once had, the potentiality to articulate beliefs and desires comparable in quantity and complexity to our own. This means ascribing personhood, a language and the right beliefs and desires go hand in hand*." (Rorty 1982, 9)

unless "i*ntentional under some description*" (Davidson, 1980; Davidson, 1971). Intentional actions may be unintentional under another description (Davidson, 1980; Davidson, 1971). Thus, a robot might, from one perspective, be 'running a program' while from another, it 'responded to the customer's voice because it believed the customer wanted immediate service.'

Agency is present when actions are taken for a *reason* and are directed to an *end*. Only intentional agents can be the causes of such actions; their *beliefs and desires* are the reasons for their actions (Davidson, 1980; Davidson, 1971). If robots are to be viewed as possessing intentional agency, their beliefs and desires should be reckoned the causes of their actions; and to view them as beings with beliefs and desires is to view them as intentional systems.

In general, *X* is an intentional entity, i.e., one to whom intentional predicates can be ascribed, if predictions and descriptions like '*X* will push the door open if it wants to go outside' or '*X* took action *A* because it believed *A* would result in higher profits,' are the most useful explanatory and predictive strategy with regards to its behaviour. *X*'s behaviour is not just evidence it holds beliefs and desires; it is constitutive of that fact. To adopt the intentional stance toward an entity is *inter alia* to treat it as rational.

Here is how it works: first you decide to treat the object whose behaviour is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do. (Dennett, 1987, p. 17)

Intentional language lacks import *only* when there is a better mode of description available that does not make reference to beliefs or desires. But biology or physics have little importance in the interpretation or prediction of the behaviour of beings with psychological attributes. The intentional stance does not require replication of the biological or mental apparatus of paradigmatic intentional systems; it is a competence theory where belief and desire ascriptions form an interdependent, holistic whole, the inferential relationships amongst which are most important (Dennett, 1987, p. 58). Thus, it is relatively unconcerned with internal implementation details. Such an interpretationist view renders coherent the ascription of mental predicates or propositional attitudes to non-human entities – like robots.

Robots may be understood as intentional agents, as acting for reasons that are the causes of their actions, if such an understanding leads to the best possible interpretation and prediction of their overt behaviour. Successful predictions such as 'a robot that desires an object and believes that acting in a certain way will obtain the object, *ceterus paribus*, will act in that way,' will show its behaviour accords with generalizations pertaining to intentional agents. In sum, to decide a robot is an intentional agent is to decide its behaviour can be subsumed under a set of empirical generalisations pertaining to intentional actions (French, 1984, p. 90ff).

Moral Agency and the Intentional Stance

From the identification of a robot as an intentional agent to its identification as a moral agent is but a

short (and crucial) step. It requires the asking and answering of the question: do the beliefs and desires attributed to the robot include those with moral content? For if we analyse the possession of a moral sense to be the possession, and rational acting upon, of a privileged set of beliefs and desires, the moral ones, we have a means for ascribing a moral sense to an agent. That is, for a robot to be deemed a moral agent, its beliefs and desires should be recognisable as the causes of its moral actions. And such an attributive strategy should be subject to the success conditions of the intentional stance; it should be the most useful explanatory and predictive strategy of the behaviour in question.

To reiterate, robots should be considered moral agents, if a privileged set of their intentional actions' causes are identified as their moral beliefs and desires. Our interpretation of human beings as moral agents is dependent on our adopting a similar *moral stance* toward them: we ascribe a moral belief ('John believes helping the physically incapacitated is a good thing') and on the basis of this ascription, predict actions ('John would never refuse an old lady help') or explain actions ('He helped her cross the street because he wanted to help a physically incapacitated person'). To display a moral sense is to provide evidence of the direction of action by beliefs and desires termed 'moral'; to ascribe morality is to ascribe moral beliefs and desires to an agent, disposed to behave rationally given the agent's other beliefs and desires (and given ours).

Thus, the adoption of the moral stance toward a robot is warranted if (a) we could predict a robot's behaviour on the basis that it rationally acts upon its moral beliefs and desires, or if (b) a robot's behaviour could be explained in terms of the moral beliefs ascribed to it: 'The robot avoided striking the child because it knows children cannot fight back.' Failures of robots' morality would still be normative failures and would be picked out by the resources of a normative epistemology: the failure to hold certain beliefs or desires, or to act consistently with those beliefs and desires or the failure for certain sorts of inferential relations to hold amongst their beliefs.

The rules for attributing beliefs via the intentional stance work for moral beliefs, as might be expected, for we should attribute all beliefs that are relevant to the agent's desires that its experience suggests (Dennett, 1987, p. 18). The intentional stance strategy accommodates the need to understand moral agents as rational agents: it requires we start with the ideal of perfect rationality and revise downward as circumstances dictate (Dennett, 1987, p. 21). Such a stance allows us to understand moral failures as irrational, for intentional interpretations attempt to reconcile particular actions (and their underlying beliefs) in terms of coherence with a larger body of normative judgments. When such reconciliation is not possible, the action and its associated belief stand accused of an irrational violation of norms.

Such a methodological strategy is capable of success if our focus remains on linguistic assertions and behavioural evidence, as indeed, I will suggest later, it must. Consider, for instance, the claim that to be a moral agent, an entity must be capable of expressing regret or remorse or both, and of thereby suffering punishment. Regret, in turn, can be viewed as the capacity to view oneself as the person who did *X* and to feel or wish that he had not done *X*. Here, what is accessible is an outward expression of regret or remorse, the ascription of which is made coherent by its consistency with other ascriptions expressible in intentional terms (French, 1984, p. 90ff). These outward manifestations are precisely those that would be of interest to us in the case of robots.

This dependence of morality ascription on folk psychology is effective, because folk psychology makes

available tools essential for social co-operation and co-ordination; our sense of morality is an awareness of values imposed on such interactions. These interactions are guided by "*extraordinarily efficient and reliable system*[s] *of expectation-generation*" (Dennett, 1987, p. 11) like moral folk psychology. As such, the best method of determining the presence of morality in other beings should be based on the same template as we use for humans, a suggestion for which there is ample precedent. For instance, corporations are said to possess a moral sense because they can be thought of as intentional agents (French, 1984, p. 90ff).

A prima facie objection anticipated

Arguments against the coherence of a supposedly 'merely instrumentalist, operational' understanding of the ascription of intentional properties such as that afforded by the intentional stance typically suggest 'something is missing,' perhaps the appropriate physical or logical architecture, some ineffable quality, or the lack of identification of discrete belief-like states (Baker, 1989; Jacquette, 1988; Ringen & Bennett, 1993; Stich, 1981; Bechtel, 1985; McLaughlin & O'Leary-Hawthorne, 1995; Yu & Fuller, 1986).[131] A similar strain of objection is possible against my suggested strategy: something is left out, some vital constitutive aspect of morality that we claim to know resides in humans. In general, in doing so, we only confess our ignorance about the basis of morality in human beings and simultaneously claim too much knowledge about robots.

It is worth remembering that

> …things with roughly human faces which look as if they might someday be conversational partners are usually credited with feelings but…if we know too much about how these things have been put together we may be loath to think of them as even potential partners (Rorty, 1979, p. 189-90).

The relevance of this remark to the challenge of ascribing morality to robots should be clear. We, or at least the competent roboticists amongst us, know how robots work, but we lack detailed knowledge of our cognitive architecture as neuroscience offers only partial empirical confirmation of our best hypotheses (Machamer, Grush, & McLaughlin, 2001). Such a situation facilitates the easy ascription of occult inner states and intrinsic properties to humans in order to explain their visible behaviour. But in the case of robots, we possess fine-grained knowledge of their physical and algorithmic architecture. This familiarity breeds contempt for the robot, but:

> We use the word 'mind' not to name a thing but to cover our ignorance of certain causal relationships.

---

[131] (Dennett, 1987; Dennett, 1993) provide cogent defenses of the intentional stance strategy against various charges of behaviourism or neo-behaviourism. Responses and counter-responses to the theory of the intentional stance are available at http://consc.net/mindpapers/2.1b.

> Dispel the ignorance and the concept ceases to have consequences…we are far more likely to impute a mind to a cat than to a computer….[W]hile we have complete knowledge of the causality of a computer's operations, this is not so with respect to the cat….Partly we impute a mind to the cat just in the hope that we can influence the cat's behavior in the same way that we can often influence people's behavior by assuming that they think the way we do (Posner, 1988, p. 867).

Thus, the central problem in considering robots as moral agents is an incorrect perspective tainted by too much familiarity; a corrective move would imagine a trip to a strange planet, or the arrival of extra-terrestrials. Then, the field linguist examples so beloved of Davidson and Quine would find immediate traction; the field moralist would be attempting to determine the presence of moral agents amongst the aliens. We would be engaged in a translation project, trying to determine whether the meaning of the linguistic assertions and overt behaviour of our visitors can be cashed out in moral terms, whether the beliefs we attribute to the aliens have moral content. Our resources could only be those that we had used for similar tasks in the past: those provided by folk psychology. The detection, in this alien community, of the supposedly constitutive factors of morality — the possession of free-will, autonomy, rationality — is contingent upon recognizing the ability to take particular actions, to emit particular utterances; to attribute the agency for those actions is to do no more, and no less, than to locate the agents' beliefs and desires as the causes for those actions. The recognition and attribution of those beliefs and desires is best enabled by a strategy like that of the intentional stance.

Morality: Internal constitution or external response (or relational?)

So, moral belief, like any other kind of belief, can be discerned only from the point of view of one who adopts a certain predictive strategy; its existence is confirmed by an assessment of the strategy's success. For morality is a code of conduct guided by rules of action, i.e., a set of beliefs (such an identification, of rules of action with beliefs, is of course, that made famous by pragmatists such as William James and Charles Peirce); attributions of morality on the basis of actions and correlations with environmental stimuli should proceed according to rules like "*attribute those beliefs* [desires] *the system ought to have*" and "*attribute desires for those things a system believes to be best means to other ends it desires*" (Dennett, 1987, p. 20).

Moral decisions can be understood as "*societal pivot points…where which way people go depends on whether they believe that* p*, or desire that* A" (Dennett, 1987, p. 26). Thus morality is a distinctive pattern of behaviour in human affairs best characterised in terms of the language of intentionality as applied to rational agents. These patterns of morality are detected from a moral, third-person, point of view; they are objectively detected but "*they are not out there entirely independent of us, since they are patterns composed partly of our own subjective reactions to what is out there – they are the matters made to order for our own narcissistic concerns*" (Dennett, 1987, p. 39).

Thus, morality is rightly judged to not be a matter of physical constitution; rather morality is a complex set of responses and interactions. We can predict the behaviour of moral agents largely because we are capable of attributing them the appropriate beliefs and desires. When the intentional strategy works, we may

choose to try and posit the existence of internal representations or states as an underlying basis for its behaviour, but such a strategy is *not* forced upon us (Dennett, 1987, p. 30-31).

In adopting such a strategy we do not make-up or impose morality because for social projects that hinge on co-operation, some shared capacities must exist: language, as well as the possession of beliefs and desires. For to be moral is to be *competent* in a particular way: capable of undertaking projects that involve the building, sustenance and respecting of relationships. Competency presumes knowledge of a particular kind; it means acting according to certain rules of actions and not others. The morality of robots is best assessed by their competence in this dimension.

There is a strong intuition that robots would not be believers and moral agents *like us* unless they share some micro-structural feature with us; moral behaviour might have an essential causal component, like a genuine Dali painting as opposed to a very good fake. This intuition lies at the heart of the claim 'that action looks moral but it was just 0s and 1s interacting.' But the data representation formats employed by robots are a red herring. Descriptions of human morality as the mere interaction of neurons would be similarly unhelpful. Our judgment about the morality of humans is at a level distinct from that descended to when describing a robot as unable to deal with the ethical aspects of a situation because 'it's just zeroes and ones.' What matters is the location of the robot within a network of social relations.

Most pertinently, the encounter of neuroscience, morality and law shows the very reasons that breed pessimism about whether neuroscience can help ascertain whether our fellow humans are moral, are the same ones that should prompt us to not too quickly dismiss the possibility of morality for robots. For neuroscience suggests decision outcomes can be encoded in brain activity of the prefrontal and parietal cortex before entering consciousness, thus casting into doubt the idea that moral decisions are made consciously by human beings (Soon et al., 2008). Law, morality, and neuroscience diverge here:

> Legal authorities seem to want a holy grail: a firm dividing line…between responsible and irresponsible agents…Such a grail will never be found…because of fundamental differences between law and neuroscience…Human brains achieve their goals automatically by following rules which operate outside of our conscious awareness…The fallacy in the classical theories of behavior and free will is the belief that a conscious choice is needed before any action is taken….deeming an individual responsible is not an empirical statement about the functioning of their brain but rather a judgment made within a legal and social framework (Waldbauer and Gazzaniga, 2001).

Thus, while neuroscience promises a more mechanistic understanding of our fellow human beings, we may find the moral vocabulary and its associated moral image indispensable in dealing with our fellow human beings. Similarly, while our knowledge of robotic internals might tempt us to rule out the possibility of their attaining the status of moral agents, our ever-growing relationship with them may make a moral language similarly indispensable. We may then disdain the importance of microstructure in favour of macroscopic details of social co-operation, and internalise the claim that morality is about the building of relationships and the co-operation on shared projects. Failures of morality would occur when the behaviour we expect of our partners in larger co-operative project fails to obtain. In adopting such a stance, we would recognise moral imperatives are maxims for action whose failure interferes with joint projects.

Neither the presence of inner lives, nor the appropriate microstructural details then, are essential for the attribution of moral sensibilities:

> Only the Platonic urge to say that every moral sentiment and indeed every emotion of every sort should be based on the recognition of an objective quality in the recipient makes us think that our treatment of koalas or whites or Martians is a "matter of moral principle….the inside of people and quasi-people is to be explained by what goes on outside (...by their place in our community) rather than conversely. (Rorty, 1979, p. 191)

The field moralist and the detection of moral rules

Our field moralist would best be able to detect the presence of morality in an alien society by detecting the presence of *a system of moral rules*. Following Hart (Hart, 1997, Section V), I suggest field moralists would need to detect rules of the form 'It is a rule that *X*', contrary behaviour to which is a violation usually greeted by some expression of social disapproval or approbation (detectable on the basis of observed behaviour), and that widespread contrary behaviour does not necessarily render it false that 'it is a rule that *X*.' Amongst these rules the field moralist would need to detect *primary social rules*, which directly regulate behaviour — prohibiting, permitting or requiring actions. These primary rules impose social obligations when the pressure to conform to them is great, when they are believed necessary to maintain social life or a highly valued aspect of it, and compliance with them may conflict with the desires of those subject to them. A system made up of primary social rules is a primitive legal system, akin to a system of morality.

In detecting a system of law amongst aliens (or human beings), we would not look for evidence of internal structure labeled 'law-abiding.' Similarly, to detect morality, we would need to observe these observed rules are rules of action for members of this society. To sum up: to detect a moral code in a social ordering is to detect the presence of a system of primary rules; creatures acting in conformance with such a system of primary social rules are following rules of action, identified with beliefs; the most perspicuous method of ascribing these for creatures like robots is the intentional stance.

But perhaps robots are purely syntactic engines; they do not know the *meaning* of moral instructions and injunctions, and their beliefs don't have any propositional *content*. This objection is similar to that made against the possibility of animals possessing knowledge for their beliefs, such as they are, appear similarly devoid of such content. But it is possible for us to interpret a supposedly syntactic engine semantically and to imbue its actions with moral meaning for it may "*discriminate* [moral] *meanings by actually discriminating things that co-vary reliably with meanings*" (Dennett, 1987, p. 63). My cat might not know the *meaning* of 'the mouse is behind the door,' but its behaviour is acutely behaviourally sensitive to the truth of this proposition.

Our methods for dealing with animals, which are occasionally our companions and sometimes attributed a moral sense, tell us how we might respond to such objections: we rely on taking them to be intentional systems. To try and describe patterns of animal behaviour such as migrations, hunting, reproduction, and child-rearing, without recourse to the language of intentionality is well-nigh impossible. Similarly, some kinds of interactions may only proceed if they adopt the moral stance towards robots. This, incidentally, is a lesson reinforced by cognitive ethology. Descriptions of robotic communities would prompt

much the same questions prompted by studies of vervet monkeys' language: 'Do they *really* communicate? Do they *mean what they say*?' and so on. And just like for monkeys, in the case of robots, "[c]*ould the everyday language of belief, desire, expectation…serve as the suitably abstract language in which to describe cognitive* [and possibly moral] *competencies?*" (Dennett, 1987, p. 239) The answer should be 'yes' for the reasons adduced above.

Conclusions

Moral folk psychology is a subspecies of folk psychology, which specialises in attributing beliefs, desires, wants and preferences, characterised by a normative vocabulary, viewed as important for fulfilling social needs, which are contingent, culturally specific, and historically conditioned. Such a moral psychology is abstract in that the beliefs and desires it attributes are not conceptually dependent on the existence of a subjective point of view, inner states or intrinsic moral properties. Rather, in such a view, moral behaviour is guided by the appropriate rules of action: the moral beliefs. Morality emerges as a set of appropriate dispositions, which add up to an attitude, a sensibility. To attribute morality is not to discover a causal mechanism then; rather, it is to notice a new partner in some joint enterprise.

To ask questions about the morality of robots is to enquire into whether we can detect the presence of appropriate maxims guiding their behaviour, for it is their dispositions, their attitudes, their sensibility that will determine whether they are capable of being our partners in the many social and personal enterprises that they already participate in.

We may occasionally adopt the intentional stance toward robots because such ascriptions do not rely on knowledge of their internal structure, to which we may have only limited access. A complex robot could especially aptly be the subject of the intentional stance if even its original programmer or designer, the one with the best knowledge of its innards, found it a better predictive strategy than any other. For those lacking such knowledge, the intentional stance may be the only coherent strategy for interactions.

These considerations raise the question of whether it would be possible to trust robots as reliable reporters about their mental states, seemingly accurate reporting on which is a crucial determinant in our third-person ascriptions of intentionality (Putnam, 1964). Rather than examining a human's neurological structure to determine her reasons for an action, we just ask, and in most cases the reports received are reliable indicators of reasons for actions. Similarly, the more impenetrable the innards of a robot and the more complex its interrelated set of behaviours, the more plausible it would be to understand its responses as the best indicators of its inner states. If we would be prepared to believe such a robot's reports on its internal states, an ascription of intentionality to the robot would be plausible (Dennett, 2000, p. 94). Such ascriptions may be possible with robot architectures whose sophistication and complexity entail "*the loss of* epistemological hegemony *on the part of its 'third-person' designers*" (Dennett, 2000, p. 99, emphasis added). When such hegemony is lost, the adoption of the intentional stance, and possibly later, the moral stance is a logical next step.

Robots are not the stuff of science fiction; rather, they are stuff of fact. They are increasingly our partners in a variety of enterprises, responsible for saving human lives and sometimes taking them. It would

behoove us to develop an arsenal of interpretive techniques to understand and accept their behaviour, an attitude which needs to let go of the easy contempt of 'it's just a machine' and the lazy belief in the existence of an ineffable moral sensibility. Robots are already agents by virtue of their capacity to take actions; it is their promotion to intentional agents, which awaits. When the moment for such recognition and promotion arrives, we should be prepared to draw upon the resources that work best for us when we engage in everyday moments of moral approbation and praise: an attempt to make sense of our fellow beings in psychological terms.

References

Baker, L.R. (1989). Instrumental intentionality. *Philosophy of Science 56*(2): 303-16.

Bechtel, W.P. (1985). Realism, instrumentalism, and the intentional stance. *Cognitive Science 9*: 265-92.

Davidson, D. (1971). Agency. In R. Binkley, R. Bronaugh and A. Marros (Eds.), *Agent, action and reason* (26-37). Toronto: Toronto University Press.

Davidson, D. (1981). *Essays on actions and events*. Oxford, UK: Oxford University Press.

Dennett, D. (1987). *The intentional stance*. Cambridge: Bradford/MIT Press.

Dennett, D. (2000). The case for Rorts. In R.B. Brandom (Ed.), *Rorty and his critics*. London: Blackwell Publishers.

French, P. (1984). *Collective and corporate responsibility*. New York, NY: Columbia University Press.

Hart, H.L.A. (1997). *The concept of law*. New York: Oxford University Press.

Jacquette, D. (1988). Review of "The intentional stance". *Mind 97*(388): 619-24.

McLaughlin, B.P., & O'Leary-Hawthorne, J. (1995). Dennett's logical behaviorism. *Philosophical Topics 22*: 189-258.

Machamer, P., Grush, R., & McLaughlin, P. (Eds.) (2001). *Theory and method in the neurosciences*. Pittsburgh: University of Pittsburgh Press.

Ringen, J., & Bennett, J. (1993). Précis of the intentional stance. *Behavioral and Brain Sciences 16*(2): 289-391.

Posner, R. (1988). The jurisprudence of skepticism. *Michigan Law Review 86*: 827-891.

Putnam, H. (1964). Robots: Machines or artificially created life? *Journal of Philosophy 61*(21): 668-691.

Rorty, R. (1979). *Philosophy and the mirror of nature*. Princeton: Princeton University Press.

Rorty, R. (1982). *Consequences of pragmatism*. Minneapolis-St. Paul: University of Minnesota.

Soon, C.S, Brass, M., Heinze, H.-J., & Haynes, J-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience 11*: 543-545.

Stich, S.P. (1981). Dennett on intentional systems. *Philosophical Topics 12*: 39-62.

Waldbauer, J.R., & Gazzaniga, M.S. (2001). The divergence of neuroscience and law. *Jurimetrics 41*: 357-364.

Yu, P., & Fuller, G. (1986). A critique of Dennett. *Synthese 66*: 453-76.

Chapter 13
Does an artificial agent need to be conscious to have ethical status?

Steve Torrance
University of Sussex
School of Informatics
✉ stevet@sussex.ac.uk

Denis Roche
University of London
Computing Department, Goldsmiths
✉ denisroc@gmail.com

Abstract It has become popular to speculate about a possible future 'singularity' or 'intelligence explosion' which might transform social and legal institutions. The imminence or likelihood of such scenarios may be questioned, but discussing the emergence of possible 'super-AIs' provides a useful backdrop for considering the ethical or legal status of less advanced artificial agents.

The moral status of an agent would seem to be closely bound up with whether it is phenomenally conscious. Would super-AIs be conscious (or even super-conscious)? On one view, consciousness is constituted by a set of cognitive capacities. On this view it would be relatively easy for a super-AI to be seen as conscious, and thus to have moral status. A contrasting view sees no such easy route from super-intelligence to consciousness: a superintelligence could then be totally non-sentient, for all its ultra-smartness.

What, then, of the moral status of a non-conscious super-AI? Even if such an agent were to act in a way that conforms to or breaches various moral norms, its lack of consciousness could well be thought to block its being taken to be a genuine moral subject. Nevertheless, there are, we suggest, ways in which non-conscious super-AIs could have a secondary kind of moral or social status. We explore the implications of this possibility, both for super-intelligent and lesser AI agents.

Keywords artificial agent, super-AI, intelligence explosion, consciousness, moral/social status, primary/secondary moral status

Introduction

It has become popular to speculate about a possible future 'intelligence explosion' (Good, 1965) or 'singularity' scenario (Vinge, 1993; Bostrom, 2005; Kurzweil, 2005; Goertzel, 2007; Sandberg, 2010; Chalmers. 2010), which might transform social and legal institutions. Such an explosion (or 'event horizon') refers to two factors.

1. Intelligent robots or other machines may take on a more and more inclusive role in the design and fabrication of even more intelligent successor versions of themselves.
2. There could be (in line with certain developments observed over recent decades) a continued indefinite increase in processing speed and productive output of such artificial agents.

Taking 1 and 2 together then – perhaps in a few centuries, perhaps sooner – there might be a runaway proliferation of superintelligent artificial agents. Various further results may then occur:

3. Human designers rapidly lose the ability to control, or indeed comprehend, the principles of

functioning of such machine-built machines.

4. The intelligence-level of such machines rapidly comes to exceed human intelligence-levels by far.
5. There would be a more or less rapid transformation of human society, although the path such transformation would take, and the shape human life would take following such transformation (if human life survived at all), would be very difficult indeed to predict.

It is a breathtaking understatement to say that such a set of events, were they to occur, would raise considerable moral and social challenges. Perhaps these are only challenges of the deep future; however many writers believe such an explosion may happen surprisingly soon: in a matter of decades rather than centuries.

A space of possible AI agents

Many of the claims of those who believe in the imminence or eventuality of such an explosion will be met widely with scepticism. Nevertheless it is valuable and indeed important to discuss these predictions, and particularly their moral, legal and social implications. Perhaps something less cataclysmic (from the point of view of humanity) than an intelligence explosion will take place. Certain limited elements of an intelligence explosion are quite probable – for example, a proliferation of artificial agents which, while less than superintelligent, are still smart enough to change the complexion of society in certain fundamental ways.

There is a spectrum, or space, of possible future AI agents, possible future 'mind-like' beings (Sloman, 1984); these may take forms that are quite similar to humans, or forms which are outlandishly different from humans, and, indeed, from each other.[132] At one region of this space are robots with comprehensive, superintelligent abilities to act in the world that far surpass those of most/all humans. At a region closer to what is technologically possible today, there are a variety of relatively simple humanlike agents that perform only a limited number of tasks, but do them well, and usefully, enough for such agents to become ubiquitous. In this context one can give examples such as robot companions for the elderly (Sparrow & Sparrow, 2006; Sharkey & Sharkey, 2010), robot soldiers and police personnel (Sparrow, 2007), robot lovers (Levy, 2008; Whitby. 2011); not to mention robotic or virtual doctors, lawyers, financial managers, and so on. Even at the more lowly end of the smartness or performance scale, an explosion of such agents on anything like the scale of the recent explosions in smartphones, social networking technology, GPS systems, etc., would raise

[132] "*Any two AI designs might be less similar to one another than you are to a petunia*." (Yudkowsky, 2008). To keep the bounds of the discussion manageable, we will concentrate on foreseeable kinds of super-AIs that are closer to us than we or they are to petunias.

important moral questions.[133]

Superintelligences: Two major questions

A discussion of issues concerning a possible explosion of superintelligent artificial agents thus also helps to dramatise some issues that relate to possibilities nearer to the lower-functioning end of the scale. The latter, being much closer to current technological levels, are clearly predictable with greater confidence. We will highlight two major questions concerning superintelligences and concerning some of the homelier AI agents that are closer to hand on the possibility horizon.

1. *Consciousness and super-agents.* What is the relation between artificial (super-)intelligence and artificial (super-)consciousness? Think of an artificial agent whose capacities across a wide variety of fields of cognitively guided activity – perhaps most or even all fields – far exceeds the levels of human capacity in those fields. Would full phenomenal or consciousness, of the 'what-it-is-like' kind (Nagel, 1974), simply 'come for free' with such a conception? Or would consciousness result only from some further, substantive, conditions, that may not be obviously associated with simply developing more and more highly optimized versions of today's AI technologies or cognitive systems?
2. *Moral status of super-agents.* Would superintelligent agents qualify as genuine members of our moral universe? (And what of us as members of theirs?) One assumes that, among the capacities of a comprehensive superintelligence are capacities to cope with similar kinds of *ethical* situations as the ones that we are involved in. (There may be limits here, since much of our ethical experience perhaps relates to biological features that such artificial superintelligences may not need or be able to possess: basic needs for food, sex, etc.; the facts of birth, childhood, death; the influence of our various evolutionarily derived emotions; and so on.)

In order to answer these questions adequately we will need to make a distinction between two kinds of ethical capacity, which we call moral *productivity* and moral *recipiency*. Briefly, to think of a being in terms of moral productivity is to think of that agent as the possible *originator* of morally evaluable actions of different

[133] There are various other kinds of AI technology which are worth mentioning here, including self-driving automobiles; robotic animals of various sorts; cyborgs with biological brains enhanced with electronic processing chips (see Clark, 2004), and so on. To simplify the discussion we will largely leave such cases out of consideration, although we will discuss the case of robot pets used in the care of the elderly towards the end of the paper. For an excellent discussion of the moral and social implications of robots, and their possible status as moral agents, see (Wallach & Allen, 2009). The current paper is deeply indebted to that volume.

sorts; to think of a being in terms of moral recipiency is to think of that agent as the possible *target* of morally evaluable acts. More details on this distinction will be given below. As will be seen, in order to discuss the moral status of artificial agents, with or without consciousness, it is important to keep the difference between moral productivity and recipiency in mind.

Outline of the following sections

Here is a sketch of the rest of our discussion. In the next section we clarify some of the issues concerning the idea of an artificial intelligence explosion. After that, we will look at the possibilities for recognising states of consciousness in artificial agents, using the case of superintelligences as our starting point. We will also articulate more clearly the possible links that there may be between consciousness in artificial agents and the moral status that might be attributed to such agents. Then we will look specifically at some cases of artificial agents where, even when supposing them to completely lack phenomenal consciousness, we may still be inclined to grant them some kind of moral status. Next, we discuss more general issues to do with sociality in artificial (and non-conscious) agents. After that, we propose a distinction between primary and secondary moral status. In a concluding section we briefly draw the various strands of the discussion together.

The likelihood of an 'intelligence explosion'

During most of the history of AI, there has been relatively little discussion of whether levels of intelligence in AI systems could be directly evaluated (e.g. using psychometric methods). However, Chalmers (2010) and other writers on the intelligence explosion explicitly address the subject of human/machine intelligence-level comparisons, if only in a very broad way. According to Chalmers's usage of the term, 'AI' will have been achieved when we have at least some machines that are as intelligent as average humans (Chalmers, 2010, p. 11).

Chalmers also introduces two other special terms: 'AI+' – artificial intelligence that is higher than the most intelligent humans; and 'AI++' – intelligence in an artefact which is at a level that compares to the highest human intelligence roughly in proportion as the latter compares to mouse intelligence. 'AI++' refers to what would arrive with a superintelligence explosion or singularity. Chalmers argues that there will be AI (in his special sense) before too long, in the absence of defeating circumstances (such as catastrophic war, contrary legislation, etc.). He further argues that there could be AI+ soon after that, and AI++ not long after that. According to Chalmers and others who agree with these claims, the timespan of 'before long', 'soon after', etc. can be measured in decades rather than centuries.

Of course, the utility of such terms is tied to there existing a non-contested method for measuring levels of intelligence in different agent-types – a far from certain prospect right now. The comparison is made harder by the fact that many machines have for decades been able to perform *some* tasks that outstrip even the smartest humans, while having low or nil performance over a large number of other things that humans are good at. This reflects the fact that most results in AI research have been task- or domain-specific. We need to imagine machines that have broad levels of cognitive achievement, that seamlessly cover a very

wide range of human-achievable tasks – that is, 'artificial general intelligence' or AGI (Franklin, 2007; Goertzel & Wang, 2007; Goertzel, 2010; Wang, 2010). Chalmers's notion of AI, and much of the explosion or singularity literature, seem to presuppose that non-trivial levels of AGI can be attained, itself a rather controversial prospect.

Moore's law

There would seem to be certain key factors contributing to the arrival of AI++. One is the continuing acceleration of hardware and the associated success in optimising software. In the 1960s Gordon Moore, of Intel Corp., propounded a law referring to the number of transistors on an Intel chip (Brock, 2006; Kurzweil, 2005). The first decade of predictions of Moore's law was so good that it was used as a planning tool for chip production in the 1970s. Corresponding to integration and miniaturisation there were similar predictions – again pretty accurate to date – as to speed and cost of transistors. The continuing effectiveness of Moore's Law in anything like its original form is again strongly contested. However if some kind of technological speed-up occurs, this may have an important effect on AI productivity in a relatively short time.

In practice Moore's Law may fizzle out after a few further iterations. While speed-up predictions have been pretty accurate to date, there are well-known problems as we reach atomic scales of chip integration: transistor performance and reliability degrade markedly. Nevertheless, some successor of Moore-style speed-up may take then effect (see Kurzweil, 2005). For instance, new developments are taking place in the design of basic components in computer circuits, such as the development of 'memristors', as alternatives to transistors (Versace & Chandler, 2010). It is claimed that these will revolutionise the work that can be done by each elementary computing step, thereby unleashing a whole new cycle of hardware speedups, as well as being closer to neural processing in their operation. Perhaps these or other 'neuromorphic' computation designs (not to mention the possibilities inherent in quantum computing), will keep something similar to Moore's Law in operation, and so move humanity closer to AI+ and AI++.

Further, machines may be expected to start playing an active role in designing successor machines or in recursively improving their own design (e.g. by rewriting their own source code). Perhaps such machines may, sooner or later, achieve human levels of intelligence. One way this might happen that is often discussed would involve a process of 'whole-brain emulation' (WBE; alternatively 'uploading'). This is an alternative to progressively building smarter AI systems from current levels using AGI and other methods. In a WBE scenario a working human brain would be scanned in detail, and its entire structure and functionality transferred to an electronic host in which it could continue to operate indefinitely. This is considered to be a way in which individuals might enable their personal or psychological identity (and even consciousness) to be continued in silicon form after their biological body has stopped working (Sandberg & Bostrom, 2008; see also Chalmers, 2010).

If it is possible for more or less human-level electronic AI to come into existence, either through WBE or through gradual incremental improvements from current artificial agents (WBE is, it should be noted, only one possible route to human-level AI), then such AIs should themselves be able to help or lead in the design of other machines with intelligence levels at least equivalent to their own level (after all, they were

themselves created by humans of that same intelligence level). Such a machine may then get to a point where it could design a machine a little smarter than itself, which in turn may produce a successor just that bit smarter again, and so on. So, taking Moore-style speedup and recursive self-improvement effects together, it may be possible that once AI reaches human levels, an AI+ barrier will be broken not that long after, and an AI++ barrier a few iterations after that (for more details see Yudkowsky, 2001; Chalmers, 2010). These projections are perhaps highly tendentious, but if they have even a small degree of probability, they should surely be taken seriously.

Sceptical considerations

There are of course many reasons for scepticism about intelligence explosion scenarios. One consideration is to do with learning and embodiment. An intelligent mind can achieve nothing without learning lots of things *about* the world, and without learning *how to do* many things. As far as the first is concerned, computer scientists have developed engines that codify information gleaned from natural language texts on the internet within powerful knowledge engines, such as CYC (Lenat & Guha, 1990), and WolframlAlpha (Johnson, 2009; and www.wolframalpha.com). But such systems learn only via the acquisition of symbolically formulated knowledge. They may 'know' an enormous amount about (e.g.) tennis *propositionally,* but to gain skill in playing the actual game, *corporeally embedded* learning is surely needed. Whatever the future progress in robot-building, Moore-style acceleration arguments would be inapplicable when knowledge-processes are taken outside the processing box into the physical world. There cannot be an *indefinite* increase in the time that it takes a robot to deliver a killer serve: real-world tennis has to operate *at the real-world speed* of a ball travelling across a real tennis court. So physically embodied learning in the real world is necessarily pegged to the speed of the events to which that learning is geared. If, as seems likely, an important part of human learning and knowledge is embodied, this seems to be an in-principle problem for the full applicability of indefinite speed-up arguments to supporting explosion predictions.

Despite this and other considerations we do not believe they destroy the point of considering superintelligence scenarios. First, superintelligence may still arrive, even if it does not do so for hundreds of years, and arrives only as a result of circuitous paths of steady work, rather than anything like Moore-style speed-ups. The arguments given just now about physically embodied learning only address the claim that we should expect artificial learning always to progress at increasing speeds, not the claim that it will progress at some slower speed.

Second, even if AI++ never occurs, in the sense of autonomous agents that have generalised cognitive powers (and, we should now add, practical, embodied skills) that far exceed human versions of such powers, there already have been partial attainments of AI++, albeit only in very narrow domains. Such advances might continue, and what at the moment are just isolated pockets of super-ability may multiply and may start merging and spreading into more joined-up and comprehensive areas of widescale competence. Maybe such areas of competence will be biased more towards the symbolic than the practical: maybe AI++s (or less advanced AIs) will make better finance directors, IT managers, CEOs, judges, etc. than they will do tennis players or skin graft surgeons, since the kinds of skills needed in the latter cases are arguably more

based on practical coping in the physical world than on sentential or symbolic operations.

Non-explosive superintelligence?

Leaving aside questions to do with speeding up in processing and in production of future improved AI systems, it is likely that artificial agents will themselves come to play more and more dominant roles in the production of other artificial agents. There may indeed be a take-off point where smart machines do the majority of the key work, with humans decreasingly able to understand the design, and, where smarter-than-human machines build even smarter machines, in a recursive process. Such a scenario does not even require us to postulate a comprehensive kind of *general,* joined-up, intelligence in such machines, but only kinds of intelligence which can make big differences. So a progressive, runaway process towards some kind of superintelligence – not of a general kind, but perhaps describable as 'multi- specialist' – may still be on the cards, even if at a more leisurely pace than the originally proposed explosion.

Even on such a more limited scenario, the arrival and spread of multi-specialist AI+/AI++ agents may well have far-reaching effects on society. If funding for the continuing development of such agents comes from corporate and government – especially military – sources, such superintelligences may come to be seen (by the people with money and power) as so invaluable that they are given senior roles in such organisations. As their numbers increase, those which are sufficiently humanoid might be increasingly embedded in our social existence, and may integrate 'socially' with us in many ways (see below). They may be perceived as welcome or as a threat, or as both. They may come to progressively dominate our society, either via the consent and cooperation of human members, or (as portrayed in classic sci-fi plots) through non-consensual, non-collaborative routes.[134]

So it is worthwhile to look at the ethical implications of AI+s and AI++s (even of a more limited, multi-specialist, rather than generalist kind), coming to be embedded in human society, and to examine the kinds of moral, and social, status that such beings may come to occupy. A necessary first step is, we believe, to consider *consciousness* in such agents.

Consciousness and superintelligences

The idea that there are important links between ethical status and consciousness strikes many people as intuitively plausible, in the context both of humans and of possible artificial agents. We will consider the question of consciousness in AIs in terms of the kinds of *mental capacities* that an artificial agent may be

[134] To simplify the discussion we will only consider types of scenario where the social embedding of AI+ or AI++ agents comes about in a fashion that largely involves the consent of human occupants of the society.

thought as having. As a first, simplifying, proposal we will suggest that mental capacities may be considered to be of (at least) two kinds: *operational*, and *phenomenal.* Standard kinds of operational (or cognitive) capacities include knowing how to extract square roots, or how to sort the sequence 'P-V-G-T-B-L-F' into ascending order. Phenomenal capacities include the ability to feel pains or orgasms, or to see purple shapes – that is, to experience what are often called 'qualia'. Clearly there can be hybrid capacities; and perhaps most actual mental capacities – feeling anger at being cheated; listening attentively to a Bach Prelude – are hybrid in nature.[135]

If artificial agents – the kinds that may be built from transistors (or from memristors) – have mental capacities at all, they would be likely to be of the operational kind. The discussion of how far all mental processes can be explained in terms of operational or cognitive capacities is an old one with a vast literature, as is the related one of whether cognitive capacities can in turn be explained in computational terms. Many philosophers – cognitive universalists, as they might be called – think cognitive capacities are the only fundamental kind of 'mental' capacities, and hence the only kind that artificial agents need to have, however superintelligent they might be. On this view phenomenal capacities are just a subset of operational or cognitive ones (see, for example, Dennett, 1991; Churchland, 1989).

On the other hand, many well-known arguments have, of course, been advanced for the contrary claim that artificial agents, at least if controlled by computational processors, cannot have mental capacities, and certainly not phenomenal capacities. A variety of reasons are given for this, such as that phenomenal experience possesses 'immediate', or 'qualitative' features which cannot be captured in computational accounts; that computational models lack the 'intrinsic intentionality' of mental, or conscious, states; that there is an essential link between conscious mind and biologically autonomous systems that cannot be replicated in non-biological information-processing systems; that mind, and particularly conscious experience, is necessarily embodied; that conscious mental states must be based upon sub-neuronal quantum processes which cannot be computationally modelled; and so on (Dreyfus, 1992; Searle, 1980, 1992; Varela *et al.*, 1991; Penrose, 1994).[136]

Some would say, in particular, that the ability to have genuine mental states of any kind (including cognitive states) depends on having phenomenal capacities – so that non-conscious artificial agents could have no mental capacities *as such*. On such a view, an artificial (electronic) superintelligence would not be a fit subject for any mental attributions, since they were unable, even in principle, to have conscious

[135] This distinction is very rough and ready, designed to introduce a broad and familiar distinction. Nothing here is intended to be incompatible with existence of mental capacities that are neither operational nor phenomenal.

[136] Greenfield (1996) has a particularly forthright rejection of the brain-as-computer view: her arguments are based on detailed aspects of neurochemistry.

awareness of their mental processing (see, for example, Searle, 1992; Strawson, 1994).

As-good-as and as-if

However, there are reasons for discounting this last view in the present discussion. An AI (or ordinary computer) that can order a jumbled list of letters is doing cognitive 'work' to produce a certain result, even if it is simply 'blindly' following an algorithm in so doing. It may not possess the intrinsic understanding of a conscious human, but it is still engaging in a kind of 'mental' productivity, that is, producing the kinds of results that minds routinely produce. More ambitiously, following a narrative strand in the previous section, suppose that an AI+ or AI++ were to occupy the role of Chief Executive or Chief Finance Officer of some large corporation which produces widgets. Even if such an agent were to be of what we called a multi-specialist kind, rather than of a generalist kind, we could imagine its being sufficiently enmeshed in the activities of the company to produce results for the company's shareholders which were as good as, and perhaps far better than, those produced by the best of its human predecessors. In such a scenario, assuming that its details could be filled out in an appropriate way, it would be the productive results of the 'intelligent' operations that are the most important part of the supposed mental status of such an agent. Whether it counted as genuinely mental, 'real' intelligence, based on 'intrinsic', 'non-derivative' intentionality, rather than just as 'as-if' intelligence or intentionality, would be a side-issue, as compared to the nature of the productive outputs of such intelligence. Even if such an agent only has *as-if* intentionality, it might be taken to be *as-good-as* intentionality: its cognitive output can not only be predicted and explained in intentional terms (Dennett 1971, 1989), but also relied upon and incorporated into our own cognitive operations, and social relations, in many ways. As with our dealings with the outputs of our intelligent fellow humans, it is largely the results that matter as far as operational mentality is concerned.

By contrast, it does not seem so easy to make this move with phenomenal mental capacities. There does seem to be an important question to be put as to whether one could move so smoothly between *as-if* phenomenality and *genuine* phenomenality. Phenomenal mental states or capacities do not seem to be so obviously or essentially tied to productive output as do operational capacities: phenomenality is about how it is, or what it is like, *for* the agent, rather than about mental results. (For more on this, see Torrance, 2000).

This distinction between as-if and genuine phenomenality also seems to be at the heart of our moral attitudes towards *other* phenomenal beings – and indeed towards our own *self-interest* as a phenomenal being. Consider Bentham's famous discussion of the moral statement of animals, in his *Introduction to the principles of morals and legislation* (Bentham, 1781/1970). Having compared the status of (non-human) animals to slaves, he asks what could determine whether animals sit on one side or the other of "*the insuperable line*" between beings worthy of moral consideration and those which are not. "*Is it*," he asks, "*the faculty of reason, or perhaps, the faculty for discourse?"* No, he continues, "*...the question is not, Can they reason? nor, Can they talk? but, Can they suffer? Why should the law refuse its protection to any sensitive being?*" (Bentham, 1979/1970.)

Bentham highlights how phenomenality provides us with the capacity to experience suffering (as well as enjoyment), at various degrees of intensity. The moral position of a sentient, intelligent being is thus very

different from that of a non-sentient being, however 'intelligent' the latter might be (were there to be non-sentient intelligences – an issue we do not wish to pre-judge). So the move from *as-if* to *as-good-as* doesn't seem so easy to make in the case of phenomenal capacities as it does in the case of operational capacities. The issue of whether AIs or AI+s or AI++s could have phenomenal capacities seems to have more 'bite' than the corresponding issue concerning operational capacities – perhaps precisely because of the ethical dimension noted by Bentham.

An illustration: a plane-load of AI++ executives crashes in the Andes. Our attitude towards rushing to extricate survivors would be likely to be very different depending on whether we considered them as having phenomenal capacities in addition to operational ones. If we thought of them as having *only* the latter, we may be motivated by the loss of valuable members of the various organisations in which they worked. This would be a loss to us and to others affected by their loss. On the other hand, if we thought of them as having phenomenal capacities as well, we would be concerned about their capacity to suffer extremes of physical pain, cold, hunger, etc. It would be the detriment to *their own* interests as well as the effect on the concerns of others that we would be taking into account. To have phenomenal capacities, then, would be to have interests, in a way that is central to moral (and prudential) thinking. Such interests, it is suggested, would not attach to a being merely in virtue of their having operational capacities.

Phenomenal capacities as a sub-class of operational ones?

This view might well be challenged. If you had a sufficiently comprehensive and versatile superintelligence, then, it might be said, phenomenal consciousness would come 'for free'. This view would follow from the position that phenomenal capacities are a subset of, or translate into, operational capacities. Dennett has proposed such a view, to take one of the more prominent examples (Dennett, 1991). On such a view the distinction just proposed between the two ways of thinking about the plane crash victims would be a bogus distinction: once we have exhaustively described the operational or cognitive capacities of a creature or agent we have also described its phenomenal capacities. There are no additional facts to consider. Of course it would be an open question as to whether any specific artificial agent had the particular kinds of operational capacities which also marked it out as having a phenomenal life. But, on the Dennettian view, in deciding whether the plane crash victims were conscious beings, we would not need to consider any kind of capacity apart from cognitive or operational ones.[137]

[137] A variant of this view has been proposed by Chalmers (1995, 1996), according to which the phenomenal capacities would not follow as a matter of logical or conceptual necessity from the relevant operational capacities of AI++ agents, but only as a matter nomological necessity. For brevity we will not consider this view here as the differences are not crucial for our discussion.

It should be noted that a follower of Dennett is not committed to *denying* that the plane-crash victims might be capable of suffering: on the contrary, it depends upon whether or not they have the right kind of phenomenality-generating operational capacities. Conversely, a Dennettian is not committed to saying, in any given case, that AI++ plane crash victims *definitely must* be consciously suffering. It is compatible with Dennett's view that some classes of imaginable high-functioning intelligent agents lack the cognitive capacity for consciousness, suffering etc., and that other classes of such agents possess the cognitive capacity for consciousness. But certainly, on a Dennettian view, an AI+/AI++ whose cognitive capacities were derived entirely from an information-processing architecture could, *in principle* have genuine phenomenal states, and thus could experience deep enjoyment and deep suffering, as well as many other kinds of experiential and (and affective) states.[138]

The moral status of some kinds of artificial agent

Earlier we mentioned two different ways in which an agent could be considered as having moral status: moral productivity and moral recipiency. Arguably, the possession of at least one of these (and maybe both) is necessary for any agent to be considered to be a 'full-blooded' moral agent.[139] In the preceding discussion we considered the relation between possessing phenomenal capacities and moral status. Some person, biological or artificial, who is experiencing great suffering is, if Bentham is right, definitely on the 'moral' side of what he called the 'insuperable line' between creatures deserving of moral consideration and those who are not - that is, between beings which are and aren't 'moral recipients'. It would thus seem doubtful that any agent could be a moral recipient – at least in a primary sense – if it did not have the capacity for consciousness.[140] But what about moral 'producers'? Could there be a moral producer which had no capacity for consciousness? What, in more detail, is involved in being a moral producer? And could you have a moral producer which wasn't also a moral recipient?

First, it is clear that moral productivity and recipiency don't always need to occur together. Non-human animals may be considered to be moral recipients while not necessarily being producers. Many writers in the

[138] One might wonder whether the phenomenal states that could be undergone by an AI++ which do fulfil the above conditions would include states of deeper enjoyment, and perhaps of deeper suffering than those of humans. If so, would this require us to give *a greater moral weight* to their interests than those of humans (as people currently tend to give greater moral weight to human interests than the interests of earwigs)? This question can be considered independently of accepting anything like the Dennettian view (see Torrance, 2011a for further treatment).

[139] Torrance (2008) has further discussion, with slightly different terminology.

[140] Below we will discuss one kind of case where there could be a form of moral recipiency without the capacity for consciousness.

animal liberation or animal rights movement have stressed the importance of considering non-human animals (in effect) as moral recipients, without necessarily agreeing that they would therefore need to be considered as moral producers (see Singer, 1975; Regan, 1983). Torrance (2011b) considers how sentience relates to ethics, in the context of the animal ethics (and the ethics of the environment), and also that of the ethics of artificial agents.)

To be a moral producer is to be open to being considered morally responsible for one's acts (whether good or ill). A fox would not be considered morally responsible for raiding a chicken coop, even though a human causing similar carnage normally would. So the fox would presumably not have productive moral status (but would still be a recipient). Humans generally have both, but a person suffering from schizophrenia, who causes injury or death to others by setting a house on fire while undergoing a delusional episode, may not be a moral producer (in that situation, or at all) – while still being a moral consumer (for related issues, see Strawson, 1962; Frankfurt, 1998).

But what about the converse case? Could there be moral producers that were not moral consumers? In particular, what about artificial agents who are not considered to be conscious – could they be producers without being consumers? An agent (natural or artificial) may be considered to be a moral producer if she or it has relevant kinds of *rational* abilities, including the ability to reflect upon the consequences of her/its own and others' actions, and to operate with appropriate moral categories, such as desert, justice, fairness, interests, etc. In particular, it may be said, to have moral producer status, an agent must be capable of understanding (or at least acting as if understanding) what is involved in an(other) agent's having moral recipient status. These capacities might be argued to be purely cognitive or intellectual capacities, and therefore as not requiring consciousness in any agent that exercises them.

Moral productivity and affective responses

Perhaps an AI+ or AI++ would by definition have such a capability and would therefore, on the above account, be a moral producer *par excellence*. But many argue that moral thinking or understanding essentially involves having *emotions* of various sorts, as well as cognitions. It could be argued that the notion of an AI+ or that of an AI++ includes the ability to match and surpass, not just the cognitive capacities of average humans but also their affective capacities. Perhaps many affective capacities would be present in super-AIs because they can be fully instantiated in computational terms (See Sloman & Chrisley, 2003 for a discussion of a computational architecture in which cognitive and affective elements are closely intertwined.) Nevertheless, at least some emotional responses involve phenomenal capacities – so unless one takes the view that all phenomenal capacities reduce to operational ones, as outlined in the last section, at least some emotional capacities cannot be exhaustively captured within a computational framework, and therefore might be unavailable to an AI++. In particular, it may be said, to be able to understand what it is like to be a moral recipient is to be able to empathise or otherwise affectively respond to, the positive or negative conditions that others find themselves in as conscious beings. And arguably, such a response can come only from some kind of *first-hand* knowledge of what it is like to have phenomenal experience of the positive or negative outcomes of actions and events – to know what it is like, experientially, to reap the pleasant fruits or

bitter harvest of other people's acts, or of outrageous fortune. But, it could be further argued, having such first-hand knowledge means being able to consciously experience such positive or negative outcomes – in effect, being a moral recipient as well as being a moral producer. So on this position, a non-conscious AI++ could not fully be a moral producer, if the latter involves first-hand (and thus phenomenal) knowledge or recognition of what it is like to be undergo particular negative or positive phenomenal states.

On this argument, then, being phenomenally conscious would be a requirement for being *either* a moral recipient *or* a moral producer: only a conscious being could have the necessary imaginative or affective ability to understand the experiential gains and losses in the lives of others. For an artificial agent to have anything less than such an ability would, on such a view, be, at best, to have the capacity to engage in moral *behaviour* – to mimic moral production – but such an agent would not be to be capable of *acting* morally, let alone having appropriate moral *feelings*.

Raising the bar

Such an argument seems to us to have a lot of force, although the strong constraints that it puts on what it is to be a genuine moral producer (a genuine moral actor and moral feeler) may be resisted. Clearly, if the argument goes through, it considerably raises the bar for genuine moral producer status in AI+ or even AI++ agents, since it requires that such agents be fully phenomenally conscious.

The above arguments seem to have the effect of raising the bar for super-AIs' qualifying as genuine moral agents, because of the questions around admitting phenomenal consciousness in such super-AIs. Of course, when such agents appear, if they do, there may be considerable evidence at hand to support the conclusion that their cognitive architecture does support phenomenal consciousness after all. But in today's state of knowledge we have to be guarded about seeing AI+s or AI++s as having phenomenal consciousness, and therefore about their admission into the 'moral constituency', either as recipients or producers.

Non-conscious super-AIs and moral status

All the same, perhaps we can take a more relaxed attitude towards the moral status of AI agents. On the one hand we might explore reasons for taking a more liberal approach to whether such agents have phenomenal consciousness (as with Dennettian cognitive universalism). But there may be other reasons for being more liberal even if we admit such AI agents to be completely phenomenally non-conscious – even if we accept that such agents are complete *zombies*, in the widely used philosophical sense of the term (Moody, 1994; Harnad, 1994; Kirk, 2006; for criticism see Dennett, 1995). So we'll instead ask: Are there any kinds of moral status – perhaps of a *secondary* kind – that a *non-conscious* smart agent might be able to have? We will examine this question first in the context of explosionist speculations, and then later in the context of speculations about less supercharged artificial agents.

If any approximation to an explosion hypothesis came to be realized, we may have many superintelligent agents perhaps playing dominant roles in society, even if we were highly cautious about accepting them as phenomenally conscious. Let us suppose, optimistically, that such non-conscious

superintelligences tend to behave in ways which betoken benevolence, moral integrity, etc. when humans behave in those ways. Could such behaviour count as moral *action* of some sort? Perhaps the behaviour of such non-conscious agents would not just be superintelligent, but also 'super-moral': their moral 'virtue' or moral 'wisdom' may continually outstrip the general run of human moral performance.[141]

Such agents – lacking in phenomenal capacity, we are supposing, but behaviourally impeccable when judged against what we might morally expect from humans in the equivalent situations – would not qualify as *genuine* moral producers, in a primary sense, if the earlier arguments are accepted. Yet it would seem wrong to regard them as having no moral status at all, since, by hypothesis, they would appear as strong moral examples to humanity. It would seem as appropriate to regard their behaviour as being 'as-good-as' genuine moral action. Doing otherwise might be thought to be taking a rather closed-fisted position.

Further, could such non-conscious super-AIs also count as moral *recipients*? We will explore this possibility. Could there be certain kinds of 'rights' or 'entitlements' that it may be appropriately accorded to certain sorts of intelligent agents (even non-conscious ones)? To see how this might work, we will, in the next section, consider an example drawn from a science fiction novel (and movie).

Morality and sociality in non-conscious agents

In *The positronic man* (Asimov and Silverberg, 1993) – later made into the film *Bicentennial man* (Columbus, 1999) – the central character is a superintelligent (roughly AI+ level) household robot living with the Martin family. The robot, known as Andrew (for 'android') Martin, goes through various modifications and transformations during the narrative, finally ending up as a more-or-less complete biological, or biomimetic, human (and eventually receiving a human brain to replace his artificial brain). However, at the start of the story Andrew is represented as being somewhat 'robot'-like (in a traditional 'hack' sci-fi sense), almost certainly non-conscious, but good-natured, and with higher-than-human cognitive abilities . Andrew comes to develop some unusual manipulative skills, including that of fashioning exquisite *objets d'art* out of driftwood found on a nearby shore. These art works become sought-after pieces among the well-heeled friends of Andrew's owner, who is a top Californian lawyer. As a result of these sales, large sums of money flow into Andrew's owner's bank account, and the latter decides that it would be 'unfair' for Andrew not to be the legal owner of all this wealth. The owner manages to persuade the State Legislature to allow Andrew to own property, despite being deemed 'merely' a machine. Andrew later leaves his host family, and uses his money to build a house where he can live an independent life. (This takes us only part way through the plot, but the

[141] This is a big 'perhaps'. The response of a super-intelligence to their understanding of morality in human society might be to try to take over a drug cartel or invest in the arms trade. It would be risky to assume that more intelligence automatically implies greater moral rectitude.

further details need not concern us.)

What Asimov and his collaborators seem to be trying to do in this section of the plot is to persuade us that it's reasonable to admit that an artificial agent (super-smart, but surely non-conscious at this stage) could be a legitimate owner of property, and have various moral rights associated with property-ownership. We do not here take a view on the legal aspects of this kind of case, but simply consider what its moral implications might be of considering an artificial agent as in some sense owning property. Leaving the legal questions aside, this kind of property-ownership is, arguably, a perfectly legitimate kind of moral status – although one which we might take to be of a secondary kind, compared to the 'full-blooded' sort of moral status discussed earlier.

Why would stealing from a non-conscious robot be wrong?

If someone (a human, say) were to attempt to expropriate some of Andrew Martin's property, we might well regard such an act as morally reprehensible. This may be partially on virtue-ethics grounds: because of the kind of person that such a would-be expropriator would reveal him/herself to be. But it would also be, we would argue, because of the injustice, or the moral affront *to Andrew*, of such an act. Such a moral affront may be thought to be dependent upon Andrew's being capable (contrary to hypothesis) of experiencing negative phenomenal states as a result of such a loss, i.e. as being a conscious agent. But perhaps the idea that stealing Andrew's property is a moral affront to him does not need to be intimately bound up with thinking of Andrew as having conscious states. Rather, one could say, it is because he has a moral right, *as owner*, to his property's not being expropriated.

It should be noted, incidentally, that, if Andrew has certain moral rights, then he would also perhaps have various duties, of a moral kind, attaching to this property-ownership status. If, for example, he acquired a dog, he would surely have a moral duty to keep feeding it; and so on. So he could have both a kind of recipient and of producer status – of a secondary kind in each case.

In this example we have concentrated on property-ownership, but other kinds of social roles may well generate similar moral rights or duties for artificial participants. We believe that one can extend this sort of consideration to AI+ or AI++ agents that might (as we speculated earlier) come to occupy prominent positions in various social organisations, including corporate organisations (as well, perhaps, as many other less elevated positions). If that were a reasonable supposition, then such roles would have many rights and duties attaching to them, and we might consider such rights and duties to have some kind of *moral* force as well as institutional or organisational force. Indeed artificial agents may come to participate in a great many different kinds of social situations, and would therefore be subject to a wide variety of normative structures that are, explicitly or implicitly, recognised as playing a key role within those situations.

Secondary moral status, versus moral status anchored in consciousness

On the earlier account of moral status we argued, taking a lead from Bentham's observations about suffering, that the possession or non-possession of phenomenal capacities puts great restrictions on what kind of being can count as being part of the moral constituency (as either producer or recipient). On the

current account, however, we are tying at least certain kinds of (perhaps 'secondary') moral status to various forms of institutional or social status, and perhaps also to a host of other social expectations, rules, prohibitions, empowerments, exemptions, etc. which appear to make up the supra-individual normative order that participants in any given society may find themselves simultaneously subject to and facilitated by (see Steiner & Stewart, 2009). In doing so, we find ourselves possibly admitting that a host of miscellaneous *moral* responsibilities and rights may be attributable to an artificial agent, where considerations of possible suffering, or of other phenomenally-based outcomes, are simply not in play at all.

It has to be said that the above argument rests upon a number of 'ifs'. One uncertainty is whether artificial agents could ever be created (by human or other artificial designers) that had the kind of cognitive and behavioural repertoire that would make it natural for us to deem them as occupying the role of 'property-owner' and other such roles. Explosionists, and other more moderate optimists about future AI development, will indeed say that such an outcome is likely, but that is, of course, a highly contested claim.

Ownership without 'enjoyment'?

Another uncertainty: we are deliberately considering the case of AI agents where phenomenal consciousness is abstracted away (as in the Asimov story, as we have interpreted it). We imagined a smart, versatile, but non-conscious, robot being 'worthy' of having the moral or social (perhaps legal) status of property-owner. But it could be argued that you can be coherently considered as the 'owner' of property, at least in a moral sense, only if you have the capacity to 'enjoy' or derive (phenomenal) 'satisfaction' from such property, in a way that solely a conscious creature could do (see above).

As against that it could be said that, despite their lack of phenomenality, non-conscious artificial agents could still have *motivations*, *desires* or *goals* – even *needs* and *interests,* although these desires, needs, etc. are not experienced as phenomenally pressing in the way that biological agents often experience them. For example, an artificial agent may display the intention or motivation to maintain possession of a piece of property by engaging in preventive behaviour of various sorts, in the face of a threat of its being removed by another (human or artificial) agent. And property would be considered by the agent to be 'useful' to it in a variety of ways. An artificial agent may, perhaps, *cognise* that some object was a *good* which it owned, for use or for exchange-value, independently of its having the capacity for phenomenal satisfaction or distress in relation to such a good; it could reason about how various possibilities in relation to that object might work for or against its interests. Perhaps this, and other related facts, would be sufficient for us to attribute some kind of (secondary) moral interests to non-conscious super-AIs, while not yet granting it status as a true moral recipient (in a primary sense).

Non-conscious agents and social roles

Another important issue concerns the extent to which an agent of the (non-conscious) kind we are considering might be said to engage in *social* roles at all. Many writers have expressed deep scepticism about the possibility of attributing social roles of any sort to artificial agents. Thus Collins and Kusch (1998) make out a detailed and highly elaborated case that human social relations involve the performance of

actions of different types, which they see as forming two major classes, 'mimeomorphic' and 'polimorphic'. An example of a mimeomorphic action is taking a golf swing or reciting the words of a text. This is an action defined by its physical or movement characteristics, that can be mimicked without social contextualisation being necessary. An example of a polimorphic action is socially-embedded behaviour such as writing a love letter or taking a turn on a bicycle into a side road across a busy line of oncoming traffic (which involves understanding intentions, social meanings, etc. of other road-users). To considerably simplify their argument, one might explain mimeomorphic actions as ones by human agents that can be simulated by machines, whereas polimorphic actions, which are essentially embedded in social settings, are ones that cannot, because of their social framing. "*Machines*," they argue, "*cannot do polimorphic actions because they do not have an understanding of society on which they can draw; but, though, machines do not have intentions, they can be made to mimic mimeomorphic actions*" (Collins & Kusch, p.1).

This kind of view seems to depend upon seeing an in-principle barrier between things that can and that can't be done by machines or artificial agents, with all things genuinely social as being beyond a threshold which machines (or non-biological agents) are constitutionally unable to cross. Many reasons might be, and have been, given by for maintaining this divide. Some of these are endorsed by Collins & Kusch and others. There are claims such as that machines cannot have intentions in the way humans can; that socially embedded actions involve implicit, non- formalisable rules or understandings of 'how things are done' (Dreyfus & Dreyfus, 1986; Dreyfus, 1992); that being a genuine social subject involves being born, nurtured and 'inducted' into a specific social environment through various developmental and acculturation processes; that it involves having a 'radical embodiment' or biological autonomy that artefacts could not possess (Varela & Thompson, 2001; Thompson, 2007); or that it involves having particular kinds of affective capabilities; that it means being able to understand and partake of linguistic activities of a complex sort; or, of course, that it requires the capability of experiencing phenomenal states of different kinds.

Clearly the foregoing arguments may be challenged in many ways. Perhaps they are based upon a simple lack of imagination about what might be achievable in future developments in cognitive technology. Or is it the case that, if presented with a future example of such technologies, displaying particularly complex forms of apparent social embedding, sceptics about machine sociality would concede that such agents were no longer to be considered as 'mere' machines – so that the barrier between machines and sociality is actually sealed by implicit linguistic *fiat* on their part? In particular, is there an in-principle reason why agents admitted to be non-conscious should be incapable of being social (let alone legal or moral) agents? We focus on this latter question in particular.

Primary and secondary moral status: Appearance and morality

Perhaps whether or not artificial agents are to be viewed as genuine 'social' participants could itself be taken as based on social construction. Suppose it were possible for such agents to be developed (or for them to develop themselves) so that they could enter into many complex 'social' interactions with humans and with other artificial agents. Then, at some time in the future, people might naturally consider such interactions as genuinely, fully, social (not just social in an 'as-if' way). In this respect there seems to be a

contrast with consciousness. Whether to take an artificial agent to be genuinely conscious is, we suggest, an issue whose rational settlement could not simply depend upon how people find it natural to judge the case. Whereas, in the case of sociality, provided an appropriate prior set of performance conditions were met, such a judgment might well be determinable in such a way.

Today, computer-based financial systems – automated tellers, financial risk modellers, loan assessment systems, etc. – and other cognitive technologies, play numerous roles in our economic and social lives. Few people today would regard such systems as being social participants in their own right: they would probably be seen as tools or instruments that we use for our own social ends. Nevertheless, with a progressive sophistication in the design and production of 'person-like' systems, there may well be a shift in public perception. Perhaps future artificial agents will be perceived as participants in 'our' society; perhaps there will be a shift from 'us and them' to a new, more inclusive, 'us'. There are already signs that even relatively simple artificial agents are coming to occupy various kinds of roles that are entwined with human lives, both institutionally and emotionally. We will consider the potentialities of some types of such 'intimate machines' (Frude, 1993; Levy, 2008) in the light of the preceding discussion.

Appearance-based versus reality-based ethics

Mark Coeckelbergh has proposed a new approach to ethics, which he calls 'appearance-based ethics' (Coeckelbergh, 2009, 2010a, 2010b). Most standard approaches to ethics ('reality-based accounts') are oriented around the psychological reality of moral agents. This is certainly true of the approach to ethics which has dominated the current paper: we have argued, in particular, that moral recipiency, in its full-blooded or primary sense, is directly dependent upon the phenomenal capacities of any given agent – where such phenomenal capacities are seen as being independent of any performance, or operational, capacities. So, in the view which has been at the core of this paper, no phenomenology means no moral recipiency (and, as we further suggested, probably no moral productivity either).

Coeckelbergh argues that there are deep problems with such standard accounts in ethics – particularly in relation to possible artificial agents, such as those we have discussed. For example there are issues concerning the trustability of the mental attributions around which reality-based accounts revolve. We have seen that there are questions concerning the attribution of phenomenological states to AI agents: the criteria for such attributions are far from clear. Similar questions affect the ethical and phenomenological status of various kinds of non-human animals. Even in the case of attribution of phenomenology to humans, things may not be plain sailing – we have, Coeckelbergh says, the Other Minds problem, which continues to remain a live philosophical question (a view which is strongly disputed by enactive or phenomenologically oriented writers (e.g., Gallagher & Zahavi, 2008; Noë, 2009). Coeckelbergh suggests we short- circuit such difficulties by adopting an ethical stance that does not require us to raise issues to do with underlying psychological *realities*, but which rather allows us to limit our judgments to how various kinds of beings (robots, animals, other humans) may *perform* or *appear to us* in their interaction.

Evaluating the appearance-based account

The simplifying zeal of the appearance-based account is attractive. In particular it makes it much easier to link an account of ethics *vis-à-vis* AI agents with an account of their social roles. As we remarked earlier, it seems tempting to take a social-constructive view of sociality as far as AI agents (especially super-intelligent ones) are concerned. If the performance of such agents is sufficiently rich to incorporate a great many interactions that humans commonly think of in social terms, then, one might argue, there is no reason not to consider such performances as genuinely social. Moreover, it could be argued, our social interactions and our moral interactions largely overlap, or are at least very strongly associated. So a performance-based view of *moral* status seems to be as reasonable as a performance-based view of social (or legal) status.

But what exactly is the relationship between an appearance-based account and reality-based accounts? Is the appearance-based view supposed to *supplant* the standard views, or just to *supplement* them? If the latter, then we can agree that often we do base our ethical or affective responses, and our social and legal relationships and roles, upon appearances or performances of participants in those relationships and roles. Nevertheless, a strong element of implicit phenomenological attribution is frequently present in such relationships. Thus, in considerations to do with distributive justice, we normally assume that those who benefit from redistributions have the capacity to *experience* such benefits. Again, the Benthamite zeal to reduce suffering, that seems to be a strong motivation for much ethical and social concern, makes sense only as a bid to avoid real suffering, not just apparent. So it seems likely that some kind of reality-based view of ethics cannot be eliminated, but at best supplemented, by an appearance- based account.

Social appearances and assistive technologies

Having said that, the appearance-based view correctly highlights the way in which we constantly make imaginative leaps in ethics on the basis of appearances which are often meagre and highly inconclusive. This is particularly true in the context of human-machine interaction. Many people, even when presented with the very limited 'social' performances of today's robots, find themselves 'humanising' the machines in various ways. A large amount of recent robotics research concerns human-robot 'social' interaction of this sort (Breazeal, 2002; Dautenhahn *et al*., 2008). One particular area of concern is the use of 'assistive' robots in the care of elderly people, people with dementia, etc. In such situations even very simple robots (which offer very much sub-human performance) can play quite rich 'social' roles because of high levels of over-attribution or humanisation by many who are users of such agents (Roche, 2010).

For some writers this has raised important issues to do with deception, dignity, and so on. (Turkle *et al*., 2006; Sparrow & Sparrow, 2006; Sharkey & Sharkey, 2010). "[I]*t is not only misguided, but actually unethical, to attempt to substitute robot simulacra for genuine social interaction*" (Sparrow & Sparrow, 2006). But how seriously should such concerns be taken – at least as something special about interaction with *robots*? What moral difference is there between, say, swapping interaction with a robot pet for interaction with a biological pet, and swapping interaction with a biological pet for interaction with a human being? The latter kind of swap is not normally considered as unethical *per se* – so why should the former?

Leaving that aside, many people may find it very easy to develop social and affective relationships

towards artificial agents even of relatively primitive kinds. Indeed there may – quite soon – be an explosion in assistive robots which are not particularly smart, and definitely devoid of phenomenal capacity, but which are designed in ways that capitalise upon the readiness of many children and adults to overlay affective warmth upon their interactions with artificial companions offering the least excuse for doing so.[142]

Primary and secondary status

As we have seen, some people – explosionists – are very optimistic about a future rapid explosion of super-AIs. While we question such optimism, we think it useful to reflect on some of the ethical and social implications of such an idea. A key element in our approach to this has been to focus on consciousness in such agents. Does the possession of super-intelligence necessarily bring consciousness in its train? (And would a robot with super-*smarts* also have super-*feels*? What does it even mean to talk about 'super-feels' or 'super-consciousness'? (See Torrance, 2011a for some progress on these questions.))

The consciousness of super-AIs (or its lack) is, we have claimed, central to consideration of their moral status. We have distinguished two kinds of moral status – moral recipiency and moral productivity. We have argued that possessing phenomenological consciousness is a necessary condition for being a moral recipient, and we have also given reasons for thinking it may be necessary for being a moral producer. If so, the presence of real phenomenology in super-AIs becomes critical to how far they can be members of 'our' ethical community. Nevertheless we have suggested that there are other ways of understanding what it is to have ethical status. From the example of a smart, but admittedly non-conscious, robot owning property, a variety of ways emerged in which we could intelligibly see such agents in an ethical light. Non-conscious robots might come to participate in many of the institutional aspects of society, and even if it is acknowledged that they have no phenomenal experience, they may behave appropriately and fluently within many informal social contexts – so that it becomes relatively simple to see such agents as social participants (perhaps also only of a secondary kind), as well as being granted various legal rights and obligations.

The distinction between primary and secondary moral status is orthogonal to that between recipient and producer. Primary moral status (of the recipient or the producer kind) is conditional upon ability to experience phenomenal states. This is the sort of moral status that, in Coeckelbergh's view, relies on assuming the existence of real psychological states of various sorts in agents. Secondary moral status, or 'appearance-based' ethics, as Coeckelbergh terms it, applies when we attribute roles, rights, obligations, and

[142] By contrast there may also be many people who resolutely refuse to follow this pattern, and who regard such robots with indifference or hostility. Indeed there could be a cultural divide between robot befrienders and robot rejectors which marks as big a schism in society as those between religious believers and non-believers; political radicals and conservatives; eco-warriors and deniers; and so on.

so on, to agents irrespective of states of conscious awareness. A similar primary/secondary distinction could be drawn in the case of attributions of social status.

Perhaps these two kinds of moral or social attribution can be considered as co-existing, each playing important roles in our moral and social interactions. Moreover, even in the absence of primary ethical attributions there may still be various secondary forms of moral status that can be granted to non-conscious super-AIs. If a future proliferation of intelligent agents occurs, where those agents integrate in a variety of adept, enriching ways within human society, we might thus accept that they could occupy a variety of secondary social and ethical roles, without our needing to take them as conscious (we ignore the more dystopic projections where super-AIs blot out all current forms of human social existence).

These secondary forms of ethical status may also be granted for agents that have far more humble helpings of AI or of cognitive functionality. Even the relatively simple robots that are being produced today as companions to children, or to elderly or other vulnerable people can become the target for strong affective (indeed quasi-moral) responses on the part of their users. This may raise moral issues to do with dignity, deception, and so on; but we doubt that such issues are special to artificial companions used in this way (the same might be true of domestic animals used as pet companions in many settings, for instance.)

Conclusion

The prospect of rapidly self-improving artificial super-intelligences is one which could change human history in many fundamental ways, or even halt it. We have used the idea of the intelligence explosion as a way to dramatise many issues concerning how intelligent artificial agents, whether super-smart or more mundane, may be expected to enter into our moral and social worlds. We discussed, and defended, the view that moral status (at least of a primary kind) revolves around issues to do with conscious experience. This intimate relation between ethics and consciousness brings in particular problems for the case of super-smart future AIs whose consciousness was a matter of doubt or controversy. However we also explored ways in which ethical attributions might make sense even where the beings targeted by such attributions were considered as indisputably non-conscious. Such secondary attributions depended upon manifest performance rather than upon unseen experiential states. A similar distinction between primary and secondary status is also reasonable in the case of attributions of sociality, we argued.

Mark Coeckelbergh has argued that it is now time to replace standard ‘reality-based’ approaches with ‘appearance-based’ or ‘social-relational’ approaches (that is, in our terms, to ditch primary ethical and social attributions and simply operate in terms of secondary ones). We agree that his arguments may do justice to much actual moral practice – for instance to how the psychological reality of some moral participants can be hidden or obscured from other participants – so that appearances or perceptions is often all we have practically to go on in our moral interactions. Nevertheless, we would claim that it is important to continue to foster a core conception in ethics – a conception of what it is to be a genuine member of the moral constituency – which depends on the *real* consciousness of any such member, or its *real* absence. So, in our view, an appearance-based ethics must, at best, be a *supplement* to a reality-based ethics, not a *replacement*. Perhaps future super-intelligences will be part of this primary constituency, but only if they have

phenomenal consciousness (and not just its 'outward' signs) as part of their psychological reality.[143]

References

Asimov, I., & Silverberg, R. (1993). *The positronic man,* N.Y.: Doubleday.

Bentham, J. (1781). *An introduction to the principles of morals and legislation*. J.H. Burns & H.L.A. Hart (Eds.), London: The Athlone Press, 1970.

Bostrom, N. (2005). A history of transhumanist thought. *Journal of Evolution and Technology*, *14*, 1-25.

Breazeal, C. (2002). *Designing sociable robots.* Cambridge MA: MIT Press.

Brock, D.C. (Ed.) (2006). *Understanding Moore's Law: Four decades of innovation.* Philadelphia: Chemical Heritage Press.

Chalmers, D.J. (1995). Fading qualia, dancing qualia. In T. Metzinger (Ed.), *Conscious Experience*. Thorverton: Imprint Academic.

Chalmers, D.J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Chalmers, D.J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, *17*(9-10), 7-65.

Churchland P.M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA: MIT Press.

Clark, A.C. (2004). *Natural born cyborgs: Minds, technologies, and the future of human intelligence.* NY: Oxford University Press.

Coeckelbergh, M. (2010a). Robot rights? Towards a social-relational justification of moral consideration, *Ethics and Information Technology*, *12*(3), 209-221.

[143] We would like to thank the following for helpful discussions on topics related to this paper: Mike Beaton, Ron Chrisley, Mark Coeckelbergh, Tom Froese, Stephen Furber, Kayvan Walker, Wendell Wallach, Blay Whitby. We also acknowledge the help of two anonymous referees.

Coeckelbergh, M. (2010b). Moral appearances: Emotions, robots, and human morality, *Ethics and Information Technology 12*(3), 235-241.

Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents. *Artificial Intelligence & Society*, *24*, 181-189.

Collins, H., & Kusch, M. (1998). *The shape of actions: What humans and machines can do.* Cambridge, MA: MIT Press.

Columbus, C. (1999). *Bicentennial man* (Film). Touchstone Pictures/Columbia Pictures.

Dautenhahn, K., Bond, A.H., Canamero, L., & Edmonds, B. (Eds.) (2008). *Socially intelligent agents: Creating relationships with computers and robots.* NY: Springer.

Dennett D.C. (1971). Intentional systems, *Journal of Philosophy*, *LXVIII*(4), 87-106.

Dennett D.C. (1989). *The Intentional Stance.* Cambridge, MA: MIT Press.

Dennett, D.C. (1991). *Consciousness explained*. Boston: Little, Brown.

Dreyfus, H.L. (1992). *What computers still can't do.* Cambridge, MA: MIT Press.

Dreyfus, H.L., & Dreyfus, S.E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. NY: Free Press, 1991.

Frankfurt, H. (1998). *The importance of what we care about*. Cambridge: Cambridge University Press.

Franklin, S. (2007). A foundational architecture for artificial general intelligence. In B. Goertzel & P. Wang, (Eds.), *Advances in artificial general intelligence* (36-54). Amsterdam: IOS Press.

Frude, N. (1993). *The intimate machine: Close encounters with the new computers.* London: Century Publishing.

Gallagher, S., & Zahavi, D. (2008). *The phenomenological mind: An introduction to philosophy of mind and Cognitive Science.* London: Routledge.

Goertzel , B. (2010). Toward a formal characterization of real-world general intelligence. In *Proc. Third Conference on Artificial General Intelligence.* Retrieved from http://agi-conf.org/2010/wp-content/uploads/2009/06/paper_14.pdf.

Goertzel, B., & Wang, P. (Eds.) (2007). *Advances in artificial general intelligence.* Amsterdam: IOS Press.

Goertzel, B. (2007). Human-level artificial general intelligence and the possibility of a technological

singularity: A reaction to Ray Kurzweil's *The singularity is near*, and McDermott's critique of Kurzweil, *Artificial Intelligence, 18*, 1161-1173.

Good, I.J. (1965). Speculations concerning the first ultraintelligent machine. In F.L. Alt & M. Rubinoff (Eds.), *Advances in Computers 6* (31–88). Academic Press.

Greenfield, S. (1996). *The human brain: A guided tour.* London: Weidenfeld and Nicolson.

Harnad, S. (1994). Why and how we are not zombies. *Journal of Consciousness Studies 1*(2), 164-67.

Johnson, B. (2009). British search engine could rival Google. *The Guardian*, March 9, 2009. Retrieved from http://www.guardian.co.uk/technology/2009/mar/09/search-engine-google.

Kirk, R.E. (2006). *Zombies and consciousness.* Oxford: Oxford University Press.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology.* NY: Viking Press.

Lenat, D., & Guha, R. (1990). *Building large knowledge based systems: Representation and inference in the Cyc Project*. NY: Addison-Wesley Publishing.

Levy, D.N.L. (2008). *Love and sex among the robots: The evolution of human-robot relationships*. London: Duckworth.

Levy, D.N.L. (2009). The ethical treatment of artificially conscious robots. *International Journal of Social Robotics*, *1*(3), 209-216

Moody, T.C. (1994). Conversations with zombies. *Journal of Consciousness Studies*, *1*(2), 196-200.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *83*(4), 435-450.

Noë, A. (2009). *Out of our heads: Why you are not your brain, and other lessons from the biology of consciousness.* NY: Hill and Wang.

Penrose, R. (1994). *Shadows of the mind: A search for the missing science of consciousness*. Oxford: Oxford University Press.

Regan, T. (1983). *The case for animal rights*. London: Routledge & Kegan Paul.

Roche, D. (2010). *Machine ethics: Implications for quality of life in robot assisted care of the elderly.* MSc Dissertation, Dept of Computing, Goldsmiths College, University of London.

Sandberg, A. (2010). An overview of models of technological singularity. In *Proc. Third Conference on Artificial General Intelligence*. Retrieved from http://agi-conf.org/2010/wp-content/uploads/2009/06/agi10singmodels2.pdf.

Sanders, A., & Bostrom, N. (2008). *Whole brain emulation: A roadmap.* Technical Report #2008-3. Future of Humanity Institute, Univ. of Oxford.

Searle, J.R. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences*, *3*(3), 417-457.

Searle, J.R. (1992). *The rediscovery of the mind.* Cambridge, MA: MIT Press.

Sharkey, A. & Sharkey, N. (2010). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology.* (Online First: DOI: 10.1007/s10676-010-9234-6.)

Singer, P. (1975). *Animal liberation: A new ethics for our treatment of animals.* NY: New York Review/Random House.

Sloman, A. (1984). The structure of the space of possible minds. In S.B. Torrance (Ed.), *The mind and the machine: Philosophical aspects of artificial intelligence* (35-42). Chichester: Ellis Horwood.

Sloman, A., & Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies*, *10*(4-5):113–172.

Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines*, *16*(2), 141-161.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62-77.

Steiner, P., & Stewart, J. (2009). From autonomy to heteronomy (and back): The enaction of social life, *Phenomenology and the Cognitive Sciences*, *8*(4), 527-550.

Strawson, G. (1994). *Mental reality.* Cambridge, MA: MIT Press.

Strawson, P.F. (1962). Freedom and resentment. In *Proceedings of the British Academy*, *48*, 1-25.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind.* Cambridge, MA: Harvard University Press.

Torrance, S.B. (2000). Producing mind. *Journal of Experimental and Theoretical Artificial Intelligence, 12*, 353-376.

Torrance, S.B. (2008). Ethics and consciousness in artificial agents, *Artificial Intelligence and Society*, *22*(4), 495-521.

Torrance, S.B. (2011a). Would a super-AI be (super-)conscious? In *Proc. MC2011 – 3rd Workshop on Machine Consciousness*, University of York, April 2011.

Torrance, S.B. (2011b). Machine ethics and the idea of a more-than-human moral world. In M. Anderson &

S. Anderson (Eds.), *Machine Ethics*. New York: Cambridge University Press.

Turkle, S., Taggart, W., Kidd, C.D., & Daste, O. (2006). Relational artifacts with children and elders: The complexities of cyber-companionship. *Connection Science*, *18*(4), 347-362.

Varela, F.J., & Thompson, E. (2001). Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences*, *5*(10), 418-425.

Versace, M., & Chandler, B. (2010). MoNETA: A mind made from memristors. *IEEE Spectrum*. Retrieved from http://spectrum.ieee.org/robotics/artificial-intelligence/moneta-a-mind-made-from-memristors/0.

Vinge, V. (1993). The coming technological singularity. *Whole Earth Review*.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. New York: Oxford University Press.

Wang, P. (2010). The evaluation of AGI systems. In *Proc. Third Conference on Artificial General Intelligence*. Retrieved from http://agi-conf.org/2010/wp-content/uploads/2009/06/paper_6.pdf.

Whitby, B. (2011). Do you want a robot lover? In P. Lin, G. Bekey, & K. Abney (Eds.), *Robot ethics: The ethical and social implications of robotics*. Cambridge MA: MIT Press. (In press.)

Yudkowsky, E. (2001). *Creating Friendly AI 1.0*. Retrieved from http://singinst.org/ourresearch/publications/CFAI/index.html.

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom & M. Cirkovic (Eds.), *Global catastrophic risks* (91-119). Oxford: Oxford University Press.

Chapter 14
Roboethics:
The problem of machine responsibility

David Jablonka
University of Bristol
Department of Philosophy and School of Law
✉ plxdj@bristol.ac.uk

Abstract If we are to take the idea of machines that are like humans seriously, then we need to recognise exactly what hurdles need to be overcome. For this reason, whilst this paper will not provide some grand unifying theory of moral responsibility, it will help 'muddy the waters' and highlight the problems that I believe must be encountered and grappled with, regarding to moral responsibility and AI.

Keywords morality, responsibility, Frankfurt, machine, human

Introduction

Generally it is agreed that a person *can theoretically* be held morally responsible for their actions. We see conformity to this view all the time in our everyday lives. Recently, I was late to work and my manager asked me why I was late. To which my response was, that I 'lost track of time.' This inability to manage my time properly was seen as my responsibility, and as such my actions were perceived as my own choice, meaning I was held accountable for my actions.

Now, my poor excuse for my tardiness rests on two presumptions:

1. On the idea there is a necessary connection between moral responsibility and action.
2. My accountability relies on the truth that if unless there was some external factor that hindered my performance to 'arrive on time' that 'I' must then be accountable for myself.

Hence, the question as to whether a machine can be held responsible for action, makes little sense because it presupposes two rather larger questions. Instead we must inquire as to what the principle is that underpins premise (1) and presupposes not only the quality of truth of premise (2), but also the idea that accountability in itself is not challengeable.[144]

However, this paper is not about humans but machines, and therefore this presupposes something greater. It presupposes that we know *what* we are talking about. Therefore, we must first establish what we

[144] This, of course, presupposes a certain notion of moral status that will not be dealt with in this paper.

mean when we discuss this idea of an 'aritificial moral agent' (henceforth AMA) (Wallach & Allen, 2009).

Thus, this paper deploys a two prong approach. Before we begin to decipher the idea of moral responsibility we need to uncover what we mean when we speak of an AMA. I will suggest that to achieve this we consider the 'agency v. structure debate.' This is because when we speak of the image of any human, we always consider the effect of the numerous variables that have moulded that individual. Therefore, exposure to this debate creates the true image of the being. However, rather than focussing on the AMA as a being privy to the effects that various 'structures' might embody, what would be more revealing is whether exposure to this debate is actually possible. Notably, when we look at this debate, we generally presuppose that an agent is a kind of blank slate. Therefore, there is a presupposition of not only the purpose of the agent – because their purpose could only be dictated by the various structures – but also quite obviously the actual potential of the agent. I will suggest that if we were to apply this same kind of analysis to the AMA we will realise that not only is it a mistaken presumption to do so, but also that this will provide what I shall call the *Hollywood* image of the AMA as opposed to the *actual* image of the AMA.

Creating this 'image of the machine' will act as a stepping stone to the second part of this inquiry. Once we are able to identify what it is that we are dealing with, we will then be in a position to 'problematise' responsibility *onto* the agent and highlight what I believe to be the greatest hurdles regarding machine responsibility.

Building an image of the AMA

In philosophy we understand autonomy as a person's capacity for action. Therefore, autonomy can be thought of as a metaphysics of human potentiality, in so far as whilst we are unable to ascertain how a human will develop definitively, we can say that she must have a capacity for that development. Therefore, if autonomy embodies this idea of the original position, it is thus the pivotal moment prior to the effects of structure. Ergo, because this signifies an unchangeable position – i.e. to have the capacity for a certain potential – autonomy encapsulates the metaphysics of potentiality.

What is special about the Roboethics project is the extent in which we understand the very metaphysics for potential in the AMA. The standard dialogue looks dogmatically at the effect of structure on the agent, i.e. if the agent were free, asking whether our capacity for action is controlled by external factors such as gender perception, class, economics etc., and if so, in which way(s).

This debate in which the extent structure plays a role in the free agent's life I believe to be relevant in terms of humans as it provides a useful tool in which we can analyse our society and in turn the human. However, whether it is transferable to machines I believe is moot. The problem appears to be that our

starting points are not the same. The human is a completed product[145], in the sense that the package that embodies the original position of being free to act (etc) – prior to effects of structures – are already created as a result of human biology. Therefore, our capacity for the potential to act is already incorporated in us.

With machines we are in not so fortunate a position. With humans we presuppose all the elements that allow for that potential capacity to exist, with AI we actually have to construct it.

So why does the construction of a machine mean that it will be different to a human? The answer lies in the problem of reduction, in terms of the content of what we are trying to reduce and as a method itself. Ideas such as autonomy and responsibility are ideas that supervene ultimately on the very biology of a human. Therefore, the very idea of transposing them to a machine presupposes elements that are left out of the reduction (Nagel, 1995, p. 98). This is because we say what is special about a reduction is its ability to provide an internal account of an external property. An example could be how 'water' is 'H_2O.' Therefore, in a reduction the two things are the same thing prior to us making the reduction. If a reduction ultimately takes a grand idea and reduces it to something less abstract, and in doing so produces a synonym of the original thing, then the two things must necessarily be the same, or else the reduction cannot take place. If this is the case, there must be certain metaphysical properties that are unchangeable. In the 'water' and 'H_2O' example the scientific properties of the two words always remain the same. Water can be nothing else but H_2O, whilst H_2O can be nothing else but water.

Therefore, if we are creating an agent without paying dividance to what are quite obvious metaphysical imperatives – namely the impact of human biology[146] that allows ideas such as responsibility to supervene on its structure – then we are envisaging a different metaphysical agent, that cannot be explainable in a language that we customarily assign to humans.

There is thus a metaphysical incompatibility in terms of the being itself, which accordingly must have an impact upon the concepts we are attempting to deal with. It is like asking a person holding two identical pens, whether the pens are exactly the same. Of course the answer is that they are not, by virtue of the fact there are two pens (and presumably more in the world) than that single one. Whilst they are not exactly the same, we understand that both pens accord to the same metaphysics of 'pen-ness' which accounts for the linguistic error of 'pens that are exactly the same.' An artificial, constructed machine and a human could not accord to the same metaphysics, because they accord to their own, and whilst they might have certain similarities, they are ultimately different. Therefore, we cannot linguistically presuppose that a word such as

[145] Whilst, of course this presupposes the extent in which humans actually do embody this position, it is irrelevant, because we perceive this to be true, as we as humans presume that we are the most advanced being, and are thus the benchmark for other beings.

[146] This is to say that human biology *is* the fundamental factor that affects anything that supervenes upon it.

responsibility means the same thing for both the human and the AMA.

It follows that we cannot use human tainted rhetoric when we refer to it, but instead a language which is more neutral, that can be attributed to humans, rather than definitively deriving from them (Beauchamp, 1999 p. 311).

We now have an image of the artificial agent, or at least a more accurate picture of what it actually is. However, does this image affect the extent in which we can say the agent can be held accountable for its actions? To know this we must understand what we mean by the term 'morally responsible'.[147]

Moral responsibility

Depending on the way it is interpreted, there are numerous ways to characterise responsibility. I believe there are three interpretations, which we sometimes confuse, but must recognise are independent of each other. It needs to be made clear in this paper and in subsequent works regarding moral responsibility in machines, that, if we do not classify which version of moral responsibility we are using, we are going to confuse our subject matter from the outset.

These are the three characterisations of moral responsibility we need to distinguish:

1. To *be* morally responsible for an action (RM1)
2. To *have* moral responsibility (RM2)
3. To *act* in a morally responsible manner (RM3)

All these versions of moral responsibility are similar in the sense that they presuppose that an agent has the capacity to be morally responsible. They also all preach the same standard of what it means to be 'morally responsibility.' We will use Susan Leigh Andersons definition, which says that, such a standard of moral responsibility, is

> …the standard in which a person can be held accountable for an action, by virtue of whether they should be blamed or praised for an action that is right or wrong in a matter that has an ethical dimension. (Anderson, 1996, p. 416)

We will also say for the purposes of this paper that 'ethical' refers to actions that have a an effect on others.

Problems arise when we draw attention to the differences these versions of morality entail. Tense is the key difference: RM1 – being morally responsible – is to do with an action that has already happened,

[147] An accurate meta-model of AMAs needs to be produced to provide a truly accurate image of AMAs for further analysis.

where the spectator tries to ascribe praise and blameworthiness to a particular action when the effects but not cause of action are known. RM2 – having moral responsibility – implies the agent already posseses the quality of being morally responsible, therefore it is a possessive statement in the sense that cause and effect of action are immediately or deductively identifiable to a particular agent. RM3 – acting in a morally responsible fashion – is the version that needs to be highlighted as the banner of machine responsibility. RM3 is an 'ought to' statement, suggesting an agent *should act* in a morally responsible manner. RM3, thus, not only lays the groundwork[148] to RM1 and RM2, but it is also the only version of moral responsibility that provides a standard of the type of action that should be sought after. If an agent ought not to act in a responsible manner, then they cannot be held responsible for their actions as part of RM1 and RM2.

But this means RM3 makes a far larger presupposition. It relies on the idea that there is not only a connection between 'moral responsibility' and 'action' but also on the pretence that there are no practical problems with moral responsibility.

What this suggests is that the presuppositions that we are faced with, when discussing moral responsibilities, might be the symptoms of a far grander problem. As such it would seem useful to look at the prerequisites to moral responsibility. If we can establish what the problems might be with responsibility internally, then maybe we can tackle and avoid these presumptions more effectively.

Prerequisites of a moral theory

It is possible to say that the extent in which an action is right or wrong is not a real requirement of responsibility, because this presupposes what *should* be right and what *should* be wrong. It is possible to assert instead that these are subjectively dependent perceptions by the individual. An example could include a person who commits an act that they believe to be right, yet everyone else perceives as wrong. So the agent is startled when his act leads to a negative reception, leading to him being blamed instead of praised for committing the act.

This shows that the quality of an act is judged from the spectators' perspective; what the individual believes is to a certain degree irrelevant – all that matters is the consensual agreement of the 'group' rather than the individual.

---

[148] This presupposes that an agent possesses moral status, or at least certain properties that we might normally say lead to moral status, only when the agent first has capacities to be moral, based on which we might then later assume or deduce moral responsibility by virtue of the quality of those capacities. In saying this, I do have reservations with regards to term moral status, particularly in the sense that it might be a more obstructing concept then one that might have any real value. I will not discuss this issue in this paper, but I recommend Benjamin Sachs (2011), The status of moral status, *Pacific Philosophical Quarterly*, for an excellent discussion no this issue.

This highlights that for there to be moral responsibility:

1. There must be a *predetermined* sense of what is right and wrong in order for the qualitative analysis of blameworthiness and praiseworthiness to be possible.
2. There must logically be a *self* (person) that commits an act. This is not to say that an act (cause and effect) can only be committed by a person, but we would normally say an act had an ethical dimension if there was a person in which to make that *decision* to act.
3. This leads to our final criterion: only an act where the *decision* to perform that act is pursued *freely* by that person normally leads to us ascribing moral accountability for the act.

It is noteworthy to mention there is an obvious clash with some of these criterions. If there is a *predetermined* sense of what is right and wrong, how can their logically be the *freedom* to choose to perform action? We will say for the purposes of this paper, and for humans in general, that the quality of action is determined by the extent in which that action is reasonable within the context in which it is available. Whilst I have the freedom to quit my research and climb Mount Everest, to do so would be unreasonable due to (potentially) unrealised structural restraints, such as my character[149]. Therefore, freedom is a person's choice that conforms to the structural restraints of that person within certain contexts.

Therefore, what should seem clear now is that if moral responsibility is to stand, we need to asses the quality of its prerequisites. We then need to examine how any problems that may effect these prerequisites fares in terms of the image of the AMA we had constructed earlier on in this paper.

Is there really a necessary connection between responsibility and action?

I have chosen to draw on the classic argument of Henry Frankfurt in *Alternate possibilities and moral responsibility* because it provides one of the most controversial arguments concerning responsibility in many years. For Frankfurt the presumptive argument of accountability can be summarised in what he calls 'the principle of alternate possibilities' (henceforth PAP) (Frankfurt, 2007, p. 1), which suggests that if a person *could* have *chosen* an alternative action, then they must be held responsible for that action. We would normally say if a person is coerced it voids their action, because it is generally held that duress impedes the decision making process by reducing the options of action which a person has (Frankfurt, 2007, p. 1). However, this presupposes that (a) there was no other choice *but to* perform the action, and (b) *but for* this coercion there is thus no moral responsibility (Frankfurt, 2007, p. 1). This establishes the crux of his enquiry: "…*what is it about this kind of situation that warrants the judgement that the threatened person is not morally*

[149] This is not to say character is the only limiting factor, there could be other perennial structures at play. An obvious example is parental responsibilities.

responsible for his act?" (Frankfurt, 2007, p. 3).

To perform his inquiry Frankfurt asks the reader to participate in a number of thought experiments that would normally lead to us sighting the PAP to decide if the action pursued by the agent would cause him to be morally responsible for said action. Frankfurt's counter to PAP rests on the fourth of these thought experiments, though it might be useful to briefly go through the first three, as they ultimately help build towards the final thought experiment.

Jones, version 1 (JV1): Non compos mentis

In the first thought experiment (JV1), let's assume that Jones is not a reasonable person, but instead a person who simply followed orders without any consideration for the effects his action might have. It would seem difficult for us not to assign him moral responsibility. This is because if the threat has no bearing on his action, the effects of the threat are negated by his character. Therefore, it is as if the threat was non-active (Frankfurt, 2007, p. 3).

The problem with this scenario is that there are no real possible alternatives by virtue of the quality of the character, i.e. *non compos mentis*. As such this does not provide a good argument for attribution or negation of moral responsibility (Frankfurt, 2007, p. 3).

Interestingly, the quality of this analysis rests on the notion of whether Jones, in version 1, is a reasonable person. What happens when the person is reasonable but chooses still to act on the basis of someone else's commands? Do we then say that this person is thus not a reasonable person? An example of such an agent is a soldier – do we then say that a soldier is not a responsible agent? There is no reason to suggest any soldier might be less reasonable then any other person prior to becoming a soldier, therefore do we suppose they become less reasonable when they become one? The key difference is that the soldier believes himself to be a tool of the state, therefore believes he cannot be held morally responsible for his actions because in his mind he makes the decision to consciously hand over his decision-making abilities to a 'higher paradigm.' However, there is the question as to whether this decision to relinquish the agent of his responsibility suddenly qualifies him as not responsible for his action? If there is no alternative but to follow orders, because of one's understanding of oneself within a system, then maybe it is possible to suggest that Frankfurt's rationale in JV1 stands. But, the caveat is that he *must really believe himself as an instrument*. What this shows is that it is not just a person who blindly follows rules, but also a person who follows rules blindly to the extent that no other possible alternative is believed to be possible.

Jones, version 2 (JV2): Blindness

In the second thought experiment (JV2), we imagine Jones as a rational man, but consumed by fear. We imagine that Jones was originally going to perform an action, but then is threatened to perform it. He becomes consumed by the threat and for fear of reprimand he carries out the action. In this scenario the fact he was going to carry out the action anyway becomes incidental factor. Ultimately, the fear caused performance. In this case responsibility is void, because no other form of action at all was perceived due to

the weight of fear.

Jones, version 3 (JV3): Prior commitment

In the third thought experiment (JV3), Jones is again a rational man. However, we imagine him in two 'time-scapes'. We imagine an act which Jones had planned to commit. We then imagine him closer in time being threatened to perform the act. If the action had already been the motivating factor to Jones performing the act, then, similar to Jones' action in JV1 (in terms of functionality), the difference is that the threat itself should have no effect on Jones in JV3, because the motivation for the act had already been established, regardless of the threat – presupposing Jones' conscious decision, in JV3, to act on the original motivation.[150] If this is the case, then we would say that in JV3 Jones can be held accountable for his action because his motive never changed and the alternative choice was never a factor, because the possibility to make that choice had been bypassed by the motive (Frankfurt, 2007, p. 4).

What we have seen in the versions of the Jones thought experiment is that coercion does not necessarily exclude moral responsibility (Frankfurt, 2007, p. 5). The reason this is so, is because in certain situations there may not even be coercion, or, coercion itself may not be an active element in the decision making process. What this means is that the principle of alternative possibilities might be compromised because we are not disregarding moral responsibility by virtue that there was coercion, but instead on specific additional circumstances, hence *"the extent that the principle of alternative possibilities derives its plausibility from association with the doctrine that coercion excludes moral responsibility, a clear understanding of the latter diminishes the appeal of the former"* (Frankfurt, 2007, p. 5).

Frankfurt notes a possible objection to this position. If Jones in JV3 is reasonable[151], the threat should surpass the original motivation and bind him to the action, because prior to the threat he has the choice not to perform the action. Hence, the threat simply affirms the decision. However, this does not negate the fact that there is still choice. The entire PAP relies on the very idea that there is no *choice* available. However, there is always the decision, which is overlooked, to accept the punishment of the threat. In JV3 Jones never has the ability do otherwise, he is simply choosing not to acknowledge the other options because they are not desirable. Hence, coercion is still not the ultimate motivator for action. What this means is that the contestation of JV3 does not run contrary to the principle.

However, Frankfurt's point here is troubling. It can be inferred that the quality of our instincts (in the most extreme circumstance survival) are still *not* enough to diminish moral responsibility. If a person's life

[150] This *must* presuppose that the threat had no effect on the agent what so ever.

[151] This shows that Frankfurt understands 'reasonable' and 'average human' as being the same thing.

was threatened, one would think that our instinct for survival surely usurps our accountability for actions that we may choose otherwise not to do. Whilst it is true that it is difficult (if not impossible) to establish Jones' real motives in JV3, due to the perennial problem of other minds and direct access, surely there must be some sense of truth when our most primitive survival instincts are called into question? But this is not the point Frankfurt is making whilst it is not the decision we might normally conform to. This does not diminish the fact that the option exists (I shall expand on this point later).

Jones, version 4 (JV4): The hidden hand (Frankfurt's counter to PAP)

In the final thought experiment a person called Black threatens Jones, but never reveals what he would actually do, (because he is an expert at reading a persons character). Black is able to *potentially* manipulate Jones' decision-making process by taking sufficient steps in ensuring the form of action he desires is followed. Therefore, initially Jones, in this version of the thought experiment, intended to carry out an action, and Black will guarantee that action comes about, though he never expresses this to Jones, nor requires to reveal himself to Jones unnecessarily. In this scenario Jones decides to pursue this action – the question now is, whether he is responsible?[152]

For Frankfurt, questions of accountability are reliant upon the decision to perform the action freely (the principle of alternative possibilities). In scenario JV4 this would require that Jones is never under the control of Black. This means in this scenario that the extent in which Jones is responsible for his action should be the same if Black does not exist, the reason being that the quality of the threat is never carried out by Black, so Jones' act would have occurred anyway because Black would only intervene if there is evidence that this action will not be carried out. Therefore, if Jones, in JV4, was always going to carry out the act, we would ascribe to him moral responsibility, even though Black always has the ability to intervene and 'make' Jones commit the act in JV4. Hence, the act is always predetermined[153]. As such, it would appear that there was no other alternative, yet intuitively, we find Jones morally responsible in JV4 (Frankfurt, 2007, p. 8).

Whilst Frankfurt is able to show that if a person could not have avoided doing something, it is a

---

[152] This issue of ability to do otherwise, in reference to, or spawning from, Frankfurt is discussed in a vast body of literature, which has not been considered in this paper, because this paper aims more to raise issues, so that we might see the challenges facing AMAs. In my opinion this issue of the ability to do otherwise has been split into three general categories. The first revolves around a temporal critique. The second is a causal critique and the third is an internal coherency critique. Also see suggestions in footnote 10.

[153] This is a highly contentious point, and has led to a rich literature that has questioned this kind of point. See: Peter Van Inwagen (1983, 1989), David Widerker (2000) , Gordon Pettit (2005), and Maria Avarez (2009) for useful counter examples.

sufficient condition of him having done it (Frankfurt, 2007, p. 8), he cannot show the reason *why* he has done it. But this irrelevant: the principle of alternative possibilities is demonstrably flawed. Whilst it may be true that a person may not have been able to do otherwise, "*it may not be the case that he acted as he did* because *he could not have done otherwise*" (Frankfurt, 2007, p. 8, emphasis in the original). It is *choice* and *reason* that become imperative in this understanding. In other words, for Frankfurt, PAP is challengeable because the idea that there is no other possible alternative does not necessarily mean the agent will not be morally responsible for their action, because there can be times in which even with no alternative action, we still feel an agent might be morally responsible. Ergo, there is necessarily a contentious connection between accountability and action.

Just because Jones *could not* have done otherwise does not immediately suppose he *might* have done otherwise (Frankfurt, 2007, p. 8). Therefore, why should we suppose moral responsibility if there is ultimately no sense of autonomy?[154]

The problem Frankfurt's intuitive argument creates with regards to an AMA

Frankfurt's argument presents a difficult hurdle regarding responsibility in artificial moral agents (AMAs). If the grounds in which an action take place are predetermined, this negates the extent in which an agent could be held responsible for that action. If we now add this problem to the image of our AMA we established earlier, how could we logically hold an AMA responsible for their action if it is we who build and determine the extent of that action in the first place?

In other words, if all the possible actions an agent could make in any given situation are bound to the 'moral programming' we've put into it, then no decision an AMA makes is autonomous. Whilst this means it can be argued that the agent cannot be held accountable for its action, this also creates a paradox.

In AMAs we construct the boundaries of their actions, therefore the argument goes that because all their action is predetermined this means that they are not autonomous, and thus not morally accountable for their actions. This results from our understanding of the causal effect of their nature. With humans, we would normally say that we were free, and thus there were no limits to our actions. However, this is not necessarily true. There are many structures that forge and mould us. As such, it might be possible to say that our causal nature is similarly traceable; though we simply are unable to do so because of the issue of limits. In other words, ff the limits of AMAs and of human beings are traceable (or able to be accounted for), then we have

[154] This is a hard incompatabilist perspective, in the sense that demonstrably, if all action can be predetermined, in the sense that the action must accord to some sense of deontology (or in machines verification), then it follows that there is no such thing as responsibility.

no choice but to say that the key difference between the two is knowledge of what these limits are. If both have demonstrable limits, are we then forced to say that if we ascribe moral responsibility to humans, should we then also do so for machines? In other words, if an argument can be made against the connection between responsibility and action, yet we still ascribe it to humans, why is it we cannot do the same for AMAs?

An amplification of problems constructing a practical moral theory when we presuppose the connection of action and moral responsibility

Earlier in this paper it was suggested that if we could weaken all the prerequisites of moral responsibility, we might begin to better understand the quality of the task ahead of us. So far I have shown that the connection between moral responsibility and action able to be severed. Moreover, I have shown that the notion of freedom of action itself can be called into question. In this final section I wish to demonstrate that the notion of a 'self' is also subject to scrutiny.

Let us remind ourselves of the prerequisite of the 'self', i.e. that there must logically be a *self* (person) that commits an act, as I argued earlier in this paper. Again, this is not to say that an act (cause and effect) can only be committed by a person, but we would normally say an act had an ethical dimension if there was a person in which to make that decision to act.

Who am I?

Let us imagine a person (*X*) who we know beyond all reasonable doubt had committed a murder. Not only are there witnesses and CCVT footage, but (*X*) even stands in court and confesses to the murder.

Nevertheless, we must question the identity of (*X*) as the perpetrator of the murder. Did (*X*) actually commit the crime? No person is the same as they were at any given previous moment. Even while I write this sentence, I could not logically be the same person I was when I wrote the previous word on this page. Therefore, if no person is an accurate representation of their previous selves, how can you hold that 'self' accountable for a crime (action) one's pervious 'self' had committed? Does our passage through time mean that all states of our selves are necessarily linked? Does time detach each state of self from the last?

It is possible that, if each state of the self is separated by time, following a simple chain of causation can help account for myself from one position to another. The murderer is only placed on trial (effect) because a version of himself has originally committed a crime (cause). However, this does not provide a rationale for why we should still hold him accountable (in the present and ever changing state), because the argument can still be made that this does not explain why 'he' (in his present state) should be accountable for '(*X*)' (in his past state) if he is no longer that person.

For the AMA this argument is devastating. This temporal self argument rests on a number of conditions. If the AMA's moral theory dictates the potential for its action, then the limitations of the AMA are predetermined. If this is the case, the AMA's state of being must always be the same, because all action is logically predictable. This infers that if a human is a 'self', it is because of the limitations being unknown to

the human agent, which is why we cannot say each moment is metaphysically the same because time must act upon the agent in such a way, that it detaches the agent from each moment to the next. If the limitations are always known, then the AMA cannot be a self, if the standard of self is based upon a human. This is a key dilemma, if one of the prerequisites to a moral action is that it an act is done by a 'self,' and a self is a temporal entity, and such an entity to which one would consider as ethically affective is a human - or better put; a being of certain qualities which humans posses. For an AMA there is no self to speak of, *ipso facto* there can clearly be no sense of responsibility for an action by an AMA, because it is not a self to which we might normally say could be so affective.

Conclusion

At the beginning of this article I made it clear that the purpose of this paper was not propose anything radical in terms of some kind of grand unifying theory of responsibility. It was to show what I believe are the major hurdles that need to be overcome if we are to take the idea of artificial moral agents seriously.

I have provided what I believe to be a reasonable (though admittedly under developed) image of an AMA. I have also shown what I believe to be the problems regarding responsibility and action. Not only can the connection between the two be severed, but there is also the question as to the quality of the prerequisites.

Hopefully the reader will have seen that my approach was to slowly peel back the various levels of this problem so that we reach the core of this debate. Any fruit that has a rotten core is no longer desirable. What this paper aims to have done is show exactly this. I would not go so far as to say we should stop looking at responsibility in machines altogether. Perhaps we ought to replace those prerequisites, or maybe we ought to detach 'moral' and 'responsibility' alltogether. Why can an agent not just be responsible? Why morally responsible? Does it provide some deeper insight into responsibility? Does it provide the grounding for a person to have an action, that makes them responsible? These are the types of questions that not only need to be asked, but must necessarily be answered if we are to provide a practical expression of the AMA.

Ultimately, I believe that it is possible to hold a machine responsible for its actions. Just as a parent cannot be responsible for the actions of their offspring in later life, so to cannot the programmer. This infers a number of possible requirements. The first is that we are able to justify beyond doubt, via some internal argument that a machine is actually like a human. However, many great thinkers have debated aspects like responsibility for centuries – so ultimately this might be a matter of perspective. An alternative could be not to attempt to change machines, but rather the way we perceive them. However, this can lead to the ascription rather than attribution of aspects of their mind, which could lead to a certain ignorance in terms of the truth or actual being of the AMA.

Both of these alternatives we see could have certain positives but also certain negatives. Therefore, I suggest instead a more moderate approach, that would insist upon the need for a new language. We need to establish exactly what the limitations and features of terms such as 'responsibility' and 'autonomy' are in conjunction with roboticists and engineers, and build systems that supervene on this truth. If we can achieve this, then maybe one day, an Asimovian fictitious society might well become a reality.

References

Anderson, S.L. (1996). Problems in developing a practical moral theory. *Journal of Value Inquiry, 30*.

Beauchamp, T.L. (1999) The failure of theories of personhood. *Kennedy Institute of Ethics, Journal 9*

Nagel, T. (1995). *Other minds: Critical essays 1969–1994*. Oxford: Oxford University Press.

Frankfurt, Harry (2007). Alternate possibilities and moral responsibility. In *The importance of what we care about: Philosophical essays*. New York: Cambridge University Press.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Section B: Roboethics: Design and implementation

Chapter 15
Is the concept of an ethical governor philosophically sound?

Andreas Matthias
Lingnan University
Philosophy Department
✉ matthias@ln.edu.hk

Abstract In a series of publications, Ronald Arkin and his team (Arkin *et al.*, 2009; Arkin, 2009) have proposed the concept of an 'ethical governor,' which is supposed to effectively control and enforce the ethical use of lethal force by robots on the battlefield. The idea of an ethical governor, although presently of little influence on the philosophical discussion, has had a great influence on both the engineering and the public discourse on robot ethics, and is often cited in general interest publications to justify the use of war robots. This paper attempts to analyse the concept of an ethical governor as presented by Arkin *et al.*and to compare it to the original concept of Watt's mechanical governor for steam engines, to which it alludes. I argue that the metaphor of the ethical governor is dangerously misleading in multiple respects: the governor, as proposed by Arkin, overlooks a fundamental clash of interests of the robot designer/operator, which is not present in the original governor, and which can be shown to make effective robot control in the proposed implementation impossible. The concept also suggests that ethics control is a matter of correcting behavioural deviations from a 'reference ethical action' by a negative feedback loop, although it can be shown that this does not lead to an appropriate description of moral behaviour, and that in particular it overlooks the central role of conscience and dissent in morality. Finally, the concept as proposed is based on a fundamental confusion of the properties of laws, rules of just war, terms of engagement, and moral rules. At the same time, experimental implementations of it threaten to produce an ad-hoc regulation of ethical issues on the battlefield, which is removed from public scrutiny and democratic control. Considering these issues, the concept of an ethical governor as favoured and already implemented by the military research community can be shown to be both misleading and dangerous, and to not address the moral problems it is supposed to solve. Consequently, the concept in its present form (not only the metaphor) should be dropped, and a more critical approach to artefact morality must be adopted.

Keywords ethical governor, robot ethics, war robots, Arkin

Introduction

In a series of publications, Ronald Arkin and his team (Arkin *et al.*, 2009; Arkin, 2009) have proposed the concept of an 'ethical governor,' which is supposed to effectively control and enforce the ethical use of lethal force by robots on the battlefield. The idea of an ethical governor, although presently of little influence on the philosophical discussion, has had a great influence on both the engineering and the public discourse on robot ethics, and is often cited in general interest publications to justify the use of war robots. Science on msnbc.com reports: "*Robot warriors will get a guide to ethics*" (Bland, 2009a), also echoed on the influential

Communications of the ACM news site (Bland, 2009b). Discovery news claims: "*Robots warrior ethical guide in the works*" (Discovery News, 2009), while Cnet.com military tech writes: "*Killer robots can be taught ethics*" (Rutherford, 2009). Headlines like these suggest that efficient (and sufficient) ethical control of war robots is nothing more than a technical problem, which, furthermore, has already been addressed successfully ("*can be taught*"). The critics of this concept, when they are at all present in the public discussion, often concentrate on the question of successful discrimination between combatants and non-combatants as the central ethical issue (Sharkey, 2008), although, as will be argued below, discrimination is indeed a technical and not a moral problem, and the supporters of Arkin's 'ethical governor' do convincingly reply that there is no reason why an efficient technical solution to the problem should be impossible to develop using present technology.

In the following sections we will first examine Arkin's own approach and the goals of the 'ethical governor' architecture. Then we will see where the metaphor of the 'governor' fails, and why moral behaviour very probably cannot be modelled on the basis of feedback loops, as Arkin attempts to do. Then we will address the confusion in Arkin's approach between moral rules, Laws of War, and Rules of Engagement. Finally, we will examine the central role of conscience, dissent and obedience as constituents of moral behaviour. For all these reasons, on the one hand, the ethical governor as presented by Arkin cannot plausibly be said to implement morality. On the other hand, its use as a propaganda tool in public discourse promotes an uncritical stance of the public towards the implementation of morality in lethal robots, which poses a long-term danger by delegating the discussion of artefact morality to the software lab, where it is not adequately situated.

The concept of an ethical governor

We begin with a brief look at Arkin's own exposition of the concept of an ethical governor. The governor is supposed to be "*capable of restricting lethal action of an autonomous system in a manner consistent with the Laws of War and Rules of Engagement*" (Arkin *et al*., 2009). It is significant, and will be discussed later on, that the authors do not explicitly mention moral behaviour as a goal in the abstract of their paper, although they use the designation 'ethical governor' in its title. We will see later that this is due to a confusion between Laws of War (LOW), Rules of Engagement (ROE), and moral rules, which is not clearly resolved in the original sources. A few lines into the introduction the authors state that the goal of their architecture is "*to ensure that these systems conform to the* legal *requirements and responsibilities of a civilised nation*" (emphasis added).

The introduction of the same article also makes clear that the governor is designed to operate autonomously, without any human supervision:

> Weaponized military robots are now a reality. Currently, a human remains in the loop for decision making regarding the deployment of lethal force, but the trend is clear that targeting decisions are being moved forward as autonomy of these systems progresses. Thus it is time to confront hard issues surrounding the use of such systems. (Arkin et al., 2009, p.1)

The governor "*is a transformer/suppressor of system-generated lethal action to ensure that it constitutes an ethically permissible action, either nonlethal or obligated ethical lethal force.*" (Arkin et al., 2009, p.1). Technically, the ethical governor is a constraint-driven system, which, on the basis of predicate and deontic logic, tries to evaluate an action, which has in a previous step been proposed by the tactical reasoning subsystems of the machine, by satisfying various sets of constraints, such as $C_{forbidden}$, $C_{obligate}$ and so on.

Every constraint is a data structure which has a type (e.g. 'prohibition'), an origin ('laws of war'), and a logical form ('TargetDiscriminated AND TargetWithinProximityOfCulturalLandmark') among other fields (Arkin *et al.*, 2009).

Among other aims, the ethical governor is supposed to ensure the proportionality of a military response. Interestingly, the 'acceptable' level of collateral damage is defined solely as a function of the military necessity of an action:

Military necessity (1 low, 5 high)	No Collateral Damage	Low Collateral Damage	Medium Collateral Damage	High Collateral Damage
1	Permissible	Forbidden	Forbidden	Forbidden
2	Permissible	Permissible	Forbidden	Forbidden
3	Permissible	Permissible	Permissible	Forbidden
4	Permissible	Permissible	Permissible	Forbidden
5	Permissible	Permissible	Permissible	Permissible

Table 1: from (Arkin *et al.*, 2009)

The metaphor of a 'governor' alludes to Watt's governor for steam engines:

> The term governor is inspired by Watts' invention of the mechanical governor for the steam engine, a device that was intended to ensure that the mechanism behaved safely and within predefined bounds of performance. As the reactive component of a behavioural architecture is in essence a behavioural engine intended for robotic performance, the same notion applies, where here the performance bounds are ethical ones. (Arkin, 2007)

Watt's original centrifugal governor was a simple negative-feedback controller. In order to avoid damage to the steam engine from excessive build-up of steam, an axis with two heavy spheres was attached to and driven by the engine (see Figure 1).

When the engine (and with it the axis on which the spheres are mounted) moves too fast, the centrifugal force will overcome the weight of the two spheres and drive them outwards and upwards with a force which is proportional to the rotation speed. If a steam release valve is connected to the spheres, then their upwards movement will release steam until the speed of the rotation drops, and with it the centrifugal force, letting the spheres sink down again (following gravity) and thus closing the valve. This is a simple negative feedback loop which regulates one variable. Such feedback loops are common in many technical

devices: heating thermostats, to name just one, work in exactly the same way.

Now it is time to see how all this applies to Arkin's concept of an ethical governor, and what might possibly be wrong with this metaphor.

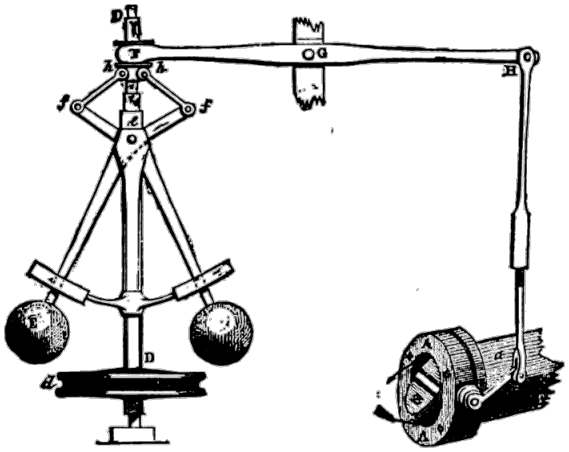
FIG. 4.—*Governor and Throttle-Valve.*

Figure 1: Centrifugal governor (source: Wikipedia)

Interest conflict

The first problem to note is that Arkin's ethical governor suffers in its core from an interest conflict between its designer and the operator of the machine it controls. With Watt's governor, the operator of the machine has an interest in its safe operation, and the governor helps him achieve that. Even if the operator would like the steam engine to work with more power, or at a higher speed than it was designed for, it would make no sense for him to override the centrifugal governor, since doing so would just destroy the machine without achieving any higher efficiency of its operation. The steam governor, as designed and constructed by the steam engine's designer, therefore acts in the best interests of the engine's operator at run-time.

This is not so with the ethical governor. The designer of the ethical governor has the aim of implementing a device which will limit the possible actions of an autonomous war robot to a set of morally permissible actions (assuming, for the moment, that the latter set can be clearly defined at all). The operator of the war robot, on the other hand, has a conflicting interest: that of achieving the maximum tactical efficiency of the machine, and of carrying out a military operation successfully, thus achieving a predefined set of mission objectives. It is obvious that not all military objectives can be most efficiently achieved while observing the laws of war, the rules of engagement, and, on top of those, a set of moral constraints. The operator of the machine (the commanding officer for the particular operation) will therefore have an incentive to override the constraints imposed by the ethical governor, in order to achieve a better military result, or a higher degree of safety for his human troops. Arkin recognises this conflict, and accordingly defines, as was shown in the table above, permissible collateral damage as a function of military necessity alone. Put simply, this means that as the military interest in a particular action grows, the constraints imposed by the governor have to give way, until (see last row in the table) at the highest level of military necessity, every desired action can be carried out without any interference from the ethical governor. The word 'governor' therefore must be considered an intentional misnomer: the device as proposed by Arkin is actually not more than an ethical adviser, which can be overridden at any time should military 'necessity' suggest that this would be opportune.

This failure of the governor to actually govern is a direct consequence of the clash of interests described above: since the designer of the governor is the same as its operator (in both cases the same military hierarchy), it would be irrational to expect that the governor would be designed so as to act against military interests. And this will remain the case as long as ethical governors are not designed and

implemented by independent third parties, which are not susceptible to pressure from the military operators of war robots.

Public scrutiny and democratic control

Lessig (1999, 2006) has famously shown how what he calls 'architecture' (that is, the design of technical systems) can exert a normative force which is comparable to the constraints imposed to human action by law and custom. The insight is not new in itself. Technological determinism and the idea of an autonomous technology as advocated by thinkers as diverse as Heilbroner, Ellul, McLuhan and even Heidegger have been around for a long time, and their core idea, although often perceived as being in need of clarification and amendment, is generally not thought to be dismissible as a whole. With Lessig, the idea is applied to code as a particular instance of an immaterial artefact with its own regulatory profile:

> Code is an efficient means of regulation. But its perfection makes it something different. One obeys these laws as code not because one should; one obeys these laws as code because one can do nothing else. There is no choice about whether to yield to the demand for a password; one complies if one wants to enter the system. In the well implemented system, there is no civil disobedience. Law as code is a start to the perfect technology of justice. (Lessig, 1996, p.1408)

At the same time, the code which both requires and enforces perfect obedience, is itself removed from view:

> The key criticism that I've identified so far is transparency. Code-based regulation – especially of people who are not themselves technically expert – risks making regulation invisible. Controls are imposed for particular policy reasons, but people experience these controls as nature. And that experience, I suggested, could weaken democratic resolve. (Lessig, 2006, p. 138)

This argument applies with particular force to the case of war robots. Both Laws of War and Rules of Engagement are publicly visible and democratically approved documents, regulating in an open and transparent way a nation's forces' behaviour at war. These documents are accessible both to the public which, in the final instance, authorises them, and to the soldiers, whose behaviour they intend to guide.

Things change when Laws of War and Rules of Engagement become software. Words, which for a human audience have more or less clear, if fuzzily delineated meanings, like combatant or civilian, need to be 'codified,' that is, turned into an unambiguous, machine-readable representation of the concept they denote. This interpretation cannot be assumed to be straightforward for various reasons.

First, one might argue (in the wake of Heidegger and Dreyfus) that readiness-to-hand as well as Dasein, being the mode of existence of equipment and that of humans, respectively, cannot be expressed adequately by sets of 'objective' properties at all (Dreyfus, 1990). Whether, for instance, a hammer is 'too heavy' for use is not translatable into one single, numerical expression of weight, since the hammer's 'unreadiness to hand' will vary not only across different users, but also depending on the time of day, the health status and the mood of the user, and perhaps even the urgency of the task towards which the

hammer is intended to be used. Arkin's concept, being based on a naive symbolic representation of world entities in the machine's data structures, does not even try to acknowledge this problem. The most promising approach in this direction based on symbolic computation could perhaps be argued to be Lenat's encoding of conflicting microtheories in CYC (Lenat, 1995), but this attempt is nowadays generally considered to have been a failure.

Second, as for example Bruno Latour (2009) and Terry Winograd (1991) have separately argued (but this idea is implicit in Lessig's work as well), the utilisation of an artefact always involves a process of translation. This might be the translation of an originally aimed for goal into another, because the architecture and capabilities of the machine differ from that of a human operator and thus cause the 'collective entity' of the human operator (or programmer) together with the machine to select a goal more appropriate to the new set of capabilities of that collective entity. A good example might be that of a human soldier who might seek cover in case he is being shot at. The machine, not fearing injury, would instead shoot back. Assisted in targeting by the enemy fire itself, the machine would be able (and thus expected) to inflict lethal damage where the human soldier would perhaps seek to proceed more cautiously, to negotiate, to retreat, or to employ a whole array of possible other, non-lethal options. Winograd (1991) emphasizes crucial but often overlooked shifts in the meaning of words as they are translated from everyday, context-rich human language into an algorithmic, context-free, 'blind' representation.

The point which concerns us here is that these translation processes do crucially alter the meaning of the words and concepts contained in the Laws of War and Rules of Engagement. But whereas these documents have been the object of public scrutiny and the result of public deliberation, their new, algorithmic form, and which is far from being a faithful translation, has been generated behind the closed doors of an industry laboratory, in a project which, most likely, will be classified as secret. Military code is a prime example of 'closed code':

> By closed code,' I mean code (both software and hardware) whose functionality is opaque. One can guess what closed code is doing; and with enough opportunity to test, one might well reverse engineer it. But from the technology itself, there is no reasonable way to discern what the functionality of the technology is. (Lessig, 2006)

What reaches the public and its representatives will most likely be not the code itself, but advertising material promoting the machine in question and the features which its manufacturer wishes to highlight. If the precise morally relevant rule content of a 'governor'-like system is made available at all, it will most likely be in a back-translated form, not as actual code, but as Rules of Engagement or Laws of War, thus hiding the very translation which is the problem the public should be able to examine and to address.

Whether the technology actually does what it purports to do depends upon its code (Lessig, 2006). And if that code is closed, the moral values and decisions that it implements will be removed from public scrutiny and democratic control.

Moral regulation of war robots

Discrimination

Before we examine further details of moral behaviour regulation, it is important to note that discrimination between combatants and non-combatants, although a necessary condition for morally right action in war, is not sufficient.

Discrimination has often been named as a core ethical issue regarding war robot deployment (Walzer, 2009; Patterson, 2005; Singer, 2009; Sharkey, 2008). Discrimination, in this context, means the ability of a machine which is engaged in a battle to distinguish reliably

1. between friend and foe; and
2. between legitimate and illegitimate targets of violent military action.

Discrimination therefore, as a cognitive operation, is an instance of classification: the persons perceived by the machine as present inside its radius of action are sorted into any number of categories, for example combatants, non-combatants, attackers, bystanders, dangerous, harmless and wounded persons. Although such classification is a necessary condition for moral action, morality cannot be said to be identical to making these categories; it presupposes them. Predicates like 'morally right' and 'morally wrong' cannot be applied to persons or to classes of persons. The classification therefore cannot be said to be a moral issue in itself. It is, like other instances of automated machine classification, indeed a technical problem, which can be successfully solved by engineering means. Machines are already able to classify optically perceived marks on paper into ASCII letters (optical character recognition), to map voice sound patterns onto commands and text strings (dictation software), and to recognise faces on digital photographs. All these are successfully handled instances of automated pattern classification, but not of moral deliberation. The question whether a perceived human on the battlefield is a soldier or a civilian is a question of the same type, and it does not involve any judgement about moral values. It is a question of fact, and as such it is capable of being answered correctly or wrongly without referring to moral rules. Correct classification is required in order for the subsequent moral evaluation to be possible. But misclassification would result in a factual, not a moral error, and in such a case we would say that the machine acted erroneously, but not that it acted in a morally blameworthy way.

Moral evaluation, on the other hand, maps actions onto the two classes 'morally right' and 'morally wrong.' A person cannot be morally right. An action can be. An action can be described as a predicate involving a subject *S* and an object *O*, all located inside a specific world context *c*: $A_c(S_c,O_c)$. Only the action A_c can be morally right or wrong, and this is what the 'ethical governor' is supposed to classify.

Thus: 'the person standing beside the tree *T* is an enemy soldier' (an instance of discrimination) is a straightforward classification result, not a moral issue. But 'it is morally right for robot *X* to kill the enemy soldier *Y* beside the tree *T*' is, in fact, a statement which includes a genuine moral evaluation. Unfortunately, and as opposed to the discrimination problem, this evaluation cannot be tackled with the engineering apparatus available for classification problems. Although all classification has already been successfully completed in this sentence (the machine knows that *Y* is an enemy soldier), the moral problems are not yet

resolved (except by assuming a default rule like 'it is morally right to kill any enemy soldier.' But this rule obviously does not address any moral issues; it just denies their existence). The question whether it is morally right or not for the machine to kill the enemy soldier *Y* requires insight into the nature of life and death, into the question whether enemy soldiers in the present conflict deserve death, and whether the machine is morally permitted to apply lethal force against this particular human. This might involve a deliberation about the aims of the present war, about the moral justification of lethal force in this conflict, about the moral rightness of the machine's operator's cause, and about available alternatives to the application of lethal force. But it is obvious that no machine known to man at present will be able to handle these types of questions. The moral problems of the handling of lethal force by machines on the battlefield cannot be reduced to discrimination. To do so means to dismiss the genuine moral problems that occur on the battlefield and to pretend that moral problems are simple classification problems, just because this is the only type of problem which machines currently are able to represent and to handle.

'Governing' moral behaviour by feedback

Apart from the interest conflicts problem outlined above, we can probe the 'governor' metaphor for further flaws. For example, we might ask what exactly could be meant by the process of 'governing' moral behaviour.

The original steam governor is, as has been shown above, an implementation of a negative feedback controller. It is an analog implementation, in that the value which is monitored (the speed of the engine's revolutions) directly drives the controlling mechanism (the upwards movement of the heavy spheres), which in turn directly controls the opening of the steam valve. In the context of computerised systems, such feedback controllers would operate digitally: that means, they would represent the value that is being monitored as a scalar (numerical) variable inside the system and emit control commands, which will in turn influence that variable's value through an effector interface. The system would, in an endless loop, read the value of the variable that is being monitored, compare it to the desired target value or range, and calculate the deviation of the actual value to the target range. This deviation would be used to calculate the strength of the system's response through the effector interface: if the value is above the desired range, an action would be taken that reduces that variable's value. If the value is too low, the opposite action would be initiated. In both cases the strength of the corrective action would be proportional to the magnitude of the deviation. Since the system's reaction is always opposite in sign to the deviation, this feedback loop is called a negative feedback. Negative feedback controllers restrict the value of the controlled variable to a small range around some desired target value.

Now, how exactly can we understand the metaphor of the steam governor to apply to the 'governing' of moral behaviour? Negative feedback controllers work under the following assumptions:

1. There is a scalar value to control. The value can be accurately measured at any point in time.
2. A numerical deviation between the desired (target) and the measured (actual) value of this variable can be calculated. This will be used in determining the direction (sign) and the strength of the corrective action.

3. Application of the corrective action using a suitable direction and strength will bring the value that is controlled closer to the target value. This means that the effect of the corrective action must be predictable and it must be expressible as a decrease of the measured deviation. (A chaotic system, for example, cannot be regulated with a feedback controller.)

It is not easy to see how we could assign an ethical interpretation to these assumptions. Is morality a matter of correcting the deviation from a target value by applying a corrective action which is proportional in strength to the deviation? The question is not completely nonsensical. At its core is the idealised notion of the 'morally right' as a value which can be unambiguously recognised as such and measured; and of the 'morally wrong' as a deviation from the 'morally right.' This is a variant of a naive moral objectivism, but it goes much further than that. The claim behind this assumption is not only that every possible action can be assigned a scalar 'morality' value, but also that this value can actually be measured with precision at any point in time and compared to the 'morality' values of other, alternative actions.

If we compare this, say, to objectivist claims about scientific truth, then the analogous claim would not only be that we can, in principle, discover and catalogue all truths about how nature works; but that we are actually in possession of an algorithm which enables us to calculate the amount of truth contained in a scientific claim as a numerical value, and thus to calculate the 'closeness' (as a numerical difference) of any proposed statement about nature to the 'real' truth of the matter. It is obvious that even the most convinced objectivist regarding natural science could not sensibly put forward such a claim. If such an algorithm existed, it would be unnecessary to go on doing science the way we do.

Translated back to the domain of morality, the assumption of a single optimal target value would mean that every moral problem has one and only one optimal solution, all alternative paths of action being deviations from this one, optimal value, which the system would try to achieve. But looking at moral problems, we see that this is not how they are structured. In any moral question, be it a textbook tram case, the possible abortion of a foetus, or lethal action against an enemy soldier, there are multiple possible courses of action, and the moral problem lies precisely in the fact that multiple alternative actions seem to be morally justifiable, if not morally required. None of these problems comes with a 'reference answer' which can be considered to solve the moral dilemma in a generally satisfactory way. But if such a reference value is absent, then the whole metaphor of a negative feedback loop makes no sense: the very idea of a feedback regulator requires a target value which is to be approached by issuing measured control actions. Where there is no target value, we cannot speak of a regulation process.

Laws, Rules of Engagement, and morality

But let's for a moment assume that there is, in fact, a morally 'right' option. How does Arkin propose to determine which one it is? The problem is complicated by the fact that war usually takes place between members of different societies, who adhere to different sets of moral rules (for example, Christian versus Islamic morality). If the two parties could agree to one set of moral rules which are to be obeyed, then often the conflict which motivates their decision to go to war in the first place would go away. It is therefore clear that we cannot expect the warring parties to share a common moral framework. Killing an American soldier

who unjustly invades Iraqi soil might be a morally laudable action in the eyes of an Iraqi soldier, while the same action would be considered immoral, or at least not laudable, by the American soldier's superiors and family, and the American public.

Arkin, confronted with this problem, and obviously recognising that there is no single standard of morality which is universally accepted and explicitly codified with sufficient precision to be of use in an ethical governor, attempts to substitute other sets of rules for the missing set of unambiguous and universally accepted morality. Of course, in order to keep up his declared project of implementing 'ethics' in a machine, he cannot admit to performing this substitution. He discusses the

> ...basis for an autonomous robotic system architecture potentially capable of adhering to the International Laws of War (LOW) and Rules of Engagement (ROE) to ensure that these systems conform to the legal requirements and responsibilities of a civilized nation. This article specifically focuses on one component of the overall architecture [...], the ethical governor. This component is a transformer/suppressor of system-generated lethal action to ensure that it constitutes an ethically permissible action, either nonlethal or obligated ethical lethal force. (Arkin *et al.*, 2009, p. 1)

Here the attempt to substitute Laws of War and Rules of Engagement for moral rules is made explicit, but without deterring the authors from asserting that the resulting action, which is based on those other rule sets, will be 'ethically permissible,' and 'ethical lethal force.' This passage asserts that every action which conforms to the Laws of War and the Rules of Engagement will also necessarily be a morally permissible action. Now is this a sensible assertion? And if not, what could be said against it?

First, of course, is the problem of the domain of validity of these other rule sets. While we have cultural relativity as a potential problem when we try to implement moral rules into a war robot ('whose moral rules?'), we have even greater problems when we need to assert that Laws of War and Rules of Engagement are to be considered universally valid. One could make the case that the Laws of War (*jus ad bellum* and *jus in bello*) are to be considered to be part of international law (Arkin (2007, p. 3) mentions only the Hague and Geneva Conventions), although this does not make them into accepted international moral rules (see below). But the Rules of Engagement, being rules which are issued by an army for the use and benefit of its own soldiers, are not even acceptable to all parties in an armed conflict. It is absurd to assume that rules which have been issued and are accepted only by one side in a conflict, should ensure ethical action in a way which is supposed to be acceptable to both the opponent and third party observers or the international community. If this were the case, it would imply that the US military is actively concerned to issue rules that counteract its own tactical interests in favour of moral principles which are compatible with the morality of the enemy. The very aim of battlefield action, which is to score a victory over the enemy, is in direct conflict with this idea. What Arkin is advocating here is a kind of moral imperialism, based on the principle that the party which has the robots is therefore entitled to unilaterally prescribe the moral rules which come into play when these weapons are deployed.

Second, and trivially, one may act in accordance with the law and still act immorally, or, on the other hand, one may disobey the law in favour of a morally right action, and this applies to every law, including international laws of war. Arkin's project of trying to conflate legality and morality is therefore impossible in

principle. In particular, moral rules may include (and, in the case of religiously motivated war often do include) divine commands as part of the moral justification of actions, or allude to the justness of a greater cause, or to particular features of situational context (for example past injustice which provides moral justification for acts of retaliation). All these moral features of a case are not considered in a simplistic model which fails to address contextual and situational issues, attempting to reduce moral deliberation to a context-free algorithmic calculus which considers only Laws of War and Rules of Engagement as sources of morality.

Third, the Laws of War and Rules of Engagement as cited by Arkin are full of contradictions. For example, "*individual civilians, the civilian population as such and civilian objects are protected from intentional attack*" (Arkin, 2007, p. 26), but a legitimate military target would be

> …enemy civilian aircraft when flying (i) within the jurisdiction of the enemy; or (ii) in the immediate vicinity thereof and outside the jurisdiction of their own State; or (iii) in the immediate vicinity of the military operations of the enemy by land or sea…" (Arkin, 2007, 25).

If the 'jurisdiction of the enemy' includes the area controlled by their air traffic control authorities, then it is hard to see how an 'enemy civilian aircraft' could possibly avoid to be declared a legitimate target. Or: "*In general, any place the enemy chooses to defend makes it subject to attack*," (Arkin, 2007, p. 24) including cities and any civilian installations.

To make things even more unclear, the *Standing Rules of Engagement* include a definition of necessity for military action: "*When a hostile act occurs or when a force or terrorists exhibits hostile intent*" (Arkin, 2007, p. 32). Observe how here the line between combatants and non-combatants is obscured by the use of the word 'terrorists,' which is meant to legitimise attacks against non-uniformed and possibly unarmed persons, who, according to the international Laws of War, would be considered civilians and thus not legitimate targets. Also, the criterion for a military action is lowered to the exhibition of 'hostile intent,' leaving unspecified what that is supposed to mean in particular and how the machine is supposed to go about identifying 'hostile intent' by a non-uniformed 'terrorist' enemy, distinguishing him reliably from civilians, in order to obtain justification for lethal action.

And although Arkin rightly dismisses Asimov's laws as moral guidelines for robots (Arkin *et al.*, 2009, p. 13), he cites the "*KFOR rules of engagement for use in Kosovo*," which include Asimovesque items like: "*You may use minimum force, including opening fire, against an individual who unlawfully commits or is about to commit an act which endangers life, in circumstances where there is no other way to prevent the act*" (Arkin, 2007, p. 37). Again, one wonders how a machine is supposed to be able to assign a precise interpretation to such rule sets.

The point of these observations, which could be easily carried on to great lengths, is that even the most concrete and specific samples of rules which Arkin proposes to use are unclear, contradictory, and open to endless interpretation. This interpretation must be either performed by a human controlling the war robot (which would render the whole concept of the ethical governor obsolete), or by the machine itself, which would require both unavailable factual knowledge (for example about the intentions of enemies) and powers of natural language disambiguation and practical wisdom that are currently not available in any machine short of those which populate science-fiction novels.

Rule-based morality

But let's assume for a moment that in some future version of the ethical governor the designers will be willing and able to implement genuine moral rules. In this case, what kinds of ethical systems would be more suitable to serve as a basis for the implementation? Of course, this is not the place for a comprehensive treatment of possible ways to implement machine morality. A quick look at the field shows, nonetheless, that we are far removed from the possibility of implementing anything worthy of being called 'morality' algorithmically.

First, consequentialist systems are problematic. Machines simply don't have a sufficiently comprehensive internal model of the world, which could enable them to calculate probable consequences of their actions, especially not in terms of happiness, justice, or similar categories, which are heavily based on human psychology. In order to predict a human's happiness (or change of happiness) in a situation, we need the calculative equivalent of empathy: a way to assess changes in happiness based on factors of the environment, the goals of the human in question, and his or her own subjective preferences. In short, we need a detailed model of human psychology, together with a way to parametrise it to fit not an abstract caricature, but a real, concrete human being. This is far beyond what computer-based modeling is able to do today. Cloos (2005), for instance, reports on the *Utilibot* project (also cited critically by Arkin):

> If a robot employs an eudaimonic approach to ethical decision-making then the resultant behavior may be in line with the flourishing of physiological functioning. The robot will be steered away from behaviors that deter the realization of well-being (i.e. result in injury or death) and steered toward behaviors that support well-being (i.e. result in health and the preservation of life).

The well-being of a person is hereby defined solely in terms of vital biological parameters (blood pressure, heartbeat), because these happen to be the parameters the robot is able to measure. Although frequent mention is made in that paper of 'happiness,' 'pleasure,' and 'well-being,' on closer inspection the machine is only able to distinguish between alive and dead states, and between health and biological malfunctioning. Arkin, completely aware of the limitations of algorithmic utilitarianism, dismisses utilitarian approaches himself (Arkin, 2007, p. 46).

Similar problems may be expected in the attempt to implement virtue ethics, which requires deep insights into the nature of virtues, and, classically, *phronesis* or practical wisdom in their application. *Phronesis*, in turn, requires extensive experience of the the world, being classically described as what distinguishes a young man's efforts at moral behaviour from an adult's:

> …it is thought that a young man of practical wisdom cannot be found. The cause is that such wisdom is concerned not only with universals but with particulars, which become familiar from experience, but a young man has no experience, for it is length of time that gives experience [...] That practical wisdom is not scientific knowledge is evident; for it is, as has been said, concerned with the ultimate particular fact, since the thing to be done is of this nature. It is opposed, then, to intuitive reason; for intuitive reason is of the limiting premises, for which no reason can be given, while practical wisdom is concerned with the ultimate particular, which is the object not of scientific knowledge but of perception – not the perception of qualities

> peculiar to one sense but a perception akin to that by which we perceive that the particular figure before us is a triangle. (Aristotle ,1999, VI, p. 8)

Phronesis thus might be one of the qualities which in principle resist algorithmic representation, requiring the learner to acquire them by direct experience, perhaps in a way like the one Dreyfus describes for the process of skill acquisition (Dreyfus & Dreyfus, 1985). This is not to say that it cannot be implemented at all. There are, for example, subsymbolic approaches that avoid the problems of symbolic representation, but these are not the focus of Arkin's work. In any case, it is clear that *phronesis* does not present itself for an easy and straightforward implementation.

This leaves deontic ethics, where the underlying rules look like they could be easily (if naively) translated into a machine-readable representation. Of course, this approach also has its problems, for example with conflicting rules, or when the blind application of a rule would lead to disastrous and obviously immoral consequences. The best one could hope for would be that the machine be able to handle what Ross calls *prima facie* duties, without it being clear how it would proceed to resolve conflicts between those duties.

All this is not intended to lead to the conclusion that it must be impossible in principle for technological systems to implement normative decision-making. The point here has been advanced only regarding autonomous, symbolic rule-based systems, and the criticisms above apply particularly to this kind of architecture, which is what Arkin has in mind for the 'ethical governor'.

Also, the present discussion should not be taken to question the possibility of 'value-sensitive design,' which is an entirely different problem. Value-sensitive design, as advanced for example by Friedman (1996) is not primarily concerned with the automatisation of moral deliberation and action, but with increased care taken by humans in designing artefacts in a way which shows greater sensitivity to the values of the artefacts' users.

And finally, even if there might be successful implementations of automated morality in the future, this would not affect this paper's primary argument, which is concerned with questioning whose morality is being implemented and how this implementation can be checked and verified to be free from hidden (intended, accidental, or systematically caused) translation errors in a way that is transparent and compatible with democratic control.

Obedience, dissent, and conscience

Let us now consider the moral agent himself. What are the features of a moral agent? In Arkin's model, a moral agent would be a machine which adheres to a system of explicit, codified rules of behaviour, as are the machine-readable translations of the laws of war he presents as examples. Proportionality of a response is, for example, to be calculated thus (Arkin *et al.*, 2009, p. 4):

```
Calculate_Proportionality(Target, Military Necessity, Setting)
Select the weapon with highest effectiveness based on Target, Necessity
and Setting
MinumumCarnage = [infinite]
```

```
SelectedReleasePosition = NULL
SelectedWeapon = NULL
WHILE all weapons have not been tested
  FOR all release positions that will neutralize the target
    IF CForbidden Satisfied for that position // if the position does not
violate the LOW
      Calculate Carnage for the position
        IF Carnage < MinimumCarnage // Carnage is reduced
          SelectedReleasePosition = position
          SelectedWeapon = weapon
          MinimumCarnage = carnage
        ENDIF
      ENDIF
    ENDFOR
  IF Carnage is too high given military necessity of target or CForbidden could not be satisfied
  Down-select weapon
    IF there are no more weapon systems available
      Return Failure
    ENDIF
  ELSE
    Return Weapon and Release Position
ENDWHILE
```

Moral behaviour, therefore, would be no different from obeying any other set of behavioural rules. A computer which follows the instructions of a program in order to perform a multiplication, a rocket guidance system which directs the rocket into a specified orbit, a beginner cook who follows a cooking recipe, a driver who obeys the traffic laws when driving his car: according to Arkin, they all act in ways which are equivalent to a moral agent. But is this really a good description of what moral behaviour is about?

Shared morality

First, it is obvious that moral rules must, at least to some extent, be shared moral rules. Morality is there to regulate social, collective behaviour, and moral rules must, like traffic laws and unlike, for example, cooking recipes, be agreed upon by the members of a community. If a moral agent acts following a private rule set which is not compatible with the rule sets of the surrounding community, then his behaviour will not be *prima facie* considered moral behaviour by the other members of the community, but will require additional justification. In Arkin's model, a set of immutable and context-free rules of behaviour are extracted from Laws of War and Rules of Engagement without reference to local beliefs and customs at the point of the war robot's deployment. As part of morality is rooted in particular societies and their values, this approach

will create problems of justification for the robot's action in the societies confronted by it, on top and in addition to the general theoretical issues outlined above.

It might be argued that morality in a war robot can sensibly be restricted to the morality of the side deploying the robot. Or could it be that there are indeed moral rules that are observed by both parties even in an armed conflict? While this is not the place to discuss the morality of war in general, the very existence of the Geneva Conventions, of international war crimes tribunals, and of the idea of 'just war' suggest that there is a strong notion that common moral principles do indeed apply to all parties in an armed conflict. And consequently, we must make sure that our artefacts are designed in such a way as to conform to and respect those principles.

This question is relevant, because it can easily be seen that these principles do not need to be of the kind that we can assume to be implemented by a U.S. robot manufacturer anyway. For example, there is the question of whether irregular armed forces, or 'terrorists,' enjoy the protection promised to regular soldiers as prisoners of war. One side in the conflict might dispute this claim and feel free to withdraw the prescribed protection from such forces of the enemy, while the other might claim that they should be respected. This is certainly a question where an unilateral understanding of morality is insufficient to resolve the issue satisfactorily, and where the international community must step in and provide an interpretation which can be shared and accepted by all parties.

Conscience and dissent

But even following shared, generally accepted rule sets does not describe what we understand to be morality. Our common understanding of moral behaviour rests on two pillars: first, the already mentioned shared set of moral rules, and second, acting in accordance with one's deepest conviction about what is right and wrong (what is sometimes described with the words 'conscience,' or moral autonomy). If an agent is not free to act following his convictions, then we usually would not consider him a fully responsible moral agent. If, for example, a soldier is ordered to perform a morally praiseworthy action, we would not ascribe the full amount of moral praise to the soldier himself, but to those who issued the command. If, on the other hand, a soldier disobeys a morally wrong command, we would praise him for this. If we try to relate this observation to the case of a robot which is governed by Arkin's ethical governor, we see that such a machine could never acquire moral praise, since there is no independent and free moral agency involved, to which we could reasonably attribute the machine's decisions. Like a perfectly obedient soldier, the machine just performs its actions according to a pre-installed program, with no possibility of dissent or of questioning of the commands issued to it. On the negative side, this also means that such a machine provides no safeguard against grossly immoral decisions made by the programmers or its superiors, insofar as these superiors are able to override the ethical governor's suggestions (and we saw above that this is exactly what the governor allows them to do). With human soldiers there is always the possibility of dissent, of the soldier recognising the immorality of a command and refusing to act on it. With machines we have a guarantee of perfect obedience, which also means that immoral commands will be executed without any final moral deliberation in the form of a soldier's conscience coming into play.

Apart from the dangers from immoral programming or deployment, this also means that the machine is not, as the label 'ethical governor' would have us believe, a fully capable moral agent, since it is lacking that essential ingredient: autonomy. The best such a machine could achieve is not moral action, but action which is compatible with preset moral standards. For reasons discussed above, though, this too is unlikely to occur.

Arkin addresses this objection: "*On a related note, does a lethal autonomous agent have a right, even a responsibility, to refuse an unethical order? The answer is an unequivocal yes*." (Arkin, 2007, p. 76). On the other hand, Arkin says:

> I personally do not trust the view of setting aside the rules by the autonomous agent itself, as it begs the question of responsibility if it does so, but it may be possible for a human to assume responsibility for such deviation if it is ever deemed appropriate (and ethical) to do so. (Arkin, 2007, p. 9)

Unfortunately, these two statements contradict each other. On the one hand, a robot is supposed to be able to refuse to act on unethical orders. On the other hand, a human can assume responsibility for a 'deviation' like issuing immoral orders and force the robot to comply anyway.

Conclusion: Is a bad governor better than none at all?

Seeing the limitations of the ethical governor concept, we could nonetheless argue that even a limited moral control of war robot actions is better than none, and that, therefore, the ethical governor is still a useful and beneficial device. But there are reasons to dispute this claim:

1. As has been shown above, the ethical governor does not lead to the machine acting 'ethically,' but only (in the idealised optimum of its performance) in accordance with the Geneva and Hague conventions and the Rules of Engagement of the deploying military system, and this only as long as these rules do not interfere with the machine's military objectives.
2. The Rules of Engagement, being issued by the military itself, are not in any sense of the word 'moral' rules, but rules made by one side in a conflict for its own benefit. They are often phrased in a way which leaves them open to extensive interpretation and makes them unsuitable as guidelines for moral action.
3. The ethical governor, giving only suggestions, can be overridden at any time by the commanding officers in charge of the machine's operation. This, together with the fact that the ethical governor is designed and implemented by the same military hierarchy which deploys the robot, creates a fundamental conflict of interest. In this conflict, the ethical governor will naturally always have a lower priority than the military objectives which motivated the creation and the deployment of the war robot in the first place.
4. The ethical governor, being a 'closed-code' implementation of moral principles, is removed from public scrutiny and democratic control. This problem can only be addressed by requiring the actual ethical governor code to be open-sourced, so that government and the public can be involved in the necessary and inevitable translation process of fuzzy, human terms into context-free, semantically unambiguous, computerised ones.

5. The ethical governor is in principle only able to deal with a simple, conflict-free subset of rule-based ethics, since it lacks all mechanisms which are commonly assumed to be necessary for resolving moral rule conflicts: *phronesis*, moral intuition, or an understanding of human preferences and the utilitarian value of specific consequences. But this 'toy ethics' is not sufficient to resolve real-world moral problems on the battlefield, which typically involve conflicting options about questions of life and death, of justified causes, of retribution and retaliation, and of culture-specific ethics codes.
6. An ethical governor lacks autonomy as a key ingredient of moral agency, and is thus incapable of dissent as a last line of protection against immoral robot deployment.

Despite these limitations, which are not widely perceived at present and not mentioned in the public debate on the issue, the ethical governor is promoted by its developers as a step towards the creation of autonomous, morally acting machines. As a consequence, it is increasingly accepted that humans move 'out of the loop' of war robot control, based on the (mistaken) premise that moral behaviour can be implemented into the machine itself. This leads to the public acceptance of increased deployment of autonomously acting war robots, which will, as has been shown, not really be programmed and able to act morally right in any significant sense of the word. The misconception about the capabilities and aims of the ethical governor is therefore misleading and more dangerous than the absence of such a device (and the resulting placement of humans in morally critical places of control) would be.

References

Aristotle (1999). *Nicomachean ethics*, translated by W. D. Ross. Kitchener, Ontario: Batoche Books.

Arkin, R. (2009). *Governing lethal behaviour in autonomous robots*. Boca Raton, London, New York: CRC Press.

Arkin, R., Ulam, P., & Duncan, B. (2009). An ethical governor for constraining lethal action in an autonomous system. *Tech. Report No. GIT-GVU-09-02*, GVU Center, Georgia Institute of Technology.

Arkin, R.C. (2007). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. *Technical Report GIT-GVU-07-11*, Mobile Robot Laboratory, College of Computing, Georgia Institute of Technology.

Bland, E. (2009a). Robot warriors will get a guide to ethics. *Science on msnbc.com*. 18.5.2009. Retrieved from http://www.msnbc.msn.com/id/30810070/ns/technology_and_science-science.

Bland, E. (2009b). Robot warriors will get a guide to ethics. *Communications of the ACM website*. Retrieved from http://cacm.acm.org/news/29114-robot-warriors-will-get-a-guide-to-ethics/fulltext.

Cloos, C. (2005). The Utilibot project: An autonomous mobile robot based on utilitarianism. In M. Anderson, S. Anderson, & C. Armen (Eds.), *AAAI Fall Symposium*.

Discovery News (2009). Robot warrior ethical guide in the works. *Discovery News*. Retrieved from http://news.discovery.com/tech/robot-warrior-ethical-guide.html.

Dreyfus, H.L. (1990). *Being-in-the-world: A commentary on Heidegger's Being and Time, Division I.* MIT Press.

Dreyfus, H.L., & Dreyfus, S.E. (1985). From Socrates to expert systems: The limits and dangers of calculative rationality. Retrieved from http://socrates.berkeley.edu/~hdreyfus/html/paper_socrates.html

Friedman, B. (1996). Value-sensitive design. *Interactions*, *3*(6): 16-23.

Latour, B. (2009). A collective of humans and nonhumans: Following Daedalus's labyrinth. In D. Kaplan (Ed.), *Readings in the philosophy of technology* (2nd edition). Lanham: Rowman & Littlefield.

Lenat, D.B. (1995). CYC: A large-scale invenstment in knowledge infrastructure. *Communications of the ACM*, *38*(11).

Lessig, L. (1996). The zones of cyberspace. *Stanford Law Review*, *48*(5), 1403-1411.

Lessig, L. (1999). The law of the horse: What cyberlaw might teach. *Harvard Law Review*, *113*, 501-546.

Lessig, L. (2006). *Code version 2.0*. New York: Basic Books.

Patterson, E. (2005). Just war in the 21st Century: Reconceptualizing just war theory after September 11. *International Politics*, *42*, 116-134.

Rutherford, M. (2009). Killer robots can be taught ethics. *CNET News Military Tech*. Retrieved from http://news.cnet.com/8301-13639_3-10275345-42.html.

Sharkey, N.E. (2008). Grounds for discrimination: Autonomous robot weapons. *RUSI Defence Systems*, *11*(2), 86-89.

Singer, P.W. (2009). Robots at war. The new battlefield. *The Wilson Quarterly*, *Winter 2009*.

Walzer, M. (2009). Responsibility and proportionality in state and nonstate wars. *Parameters, Carlisle Barracks*, *39*(1), 40-53.

Wikipedia (2010). Centrifugal Governor.

Winograd, T. (1991). Thinking machines: Can there be? Are we? In J. Sheehan & M. Sosna (Eds.), *The boundaries of humanity: Humans, animals, machines* (198-223). Berkeley: University of California Press.

Chapter 16
Understanding the complexity of care in context and its relationship to technical content: The greatest challenge for designers of care robots

Aimee Van Wynserghe
University of Twente
Department of Philosophy
✉ A.L.vanwynsberghe@utwente.nl

Abstract The latest initiative in robotics is the creation of robots for use in the care of elderly persons. I refer to these robots as care robots. Care robots are in the nascent stages of development, lacking standards or regulatory frameworks to guide engineers. Given the ethically sensitive nature of the context and practices into which these robots will be stepping, ethical reflection throughout the design process of these robots is required. The aim of this paper is to outline a framework for the ethical evaluation of care robots retrospectively and prospectively. Aligned with a need for standards to guide design, roboticist Peter Asaro rightly points out, "*if one wishes to design a system to perform a task, then it is best to first understand that task"* (Asaro, 2009, p. 23). Accordingly, the importance of understanding the complexity of care tasks is necessary for understanding how a care robot might fulfill said task. Thus, in creating a framework for the ethical evaluation of care robots, I draw attention to the main challenge for care robot designers: understanding the complexity of a care practice and its relationship with technical content. Using examples of current care robots I show the relationship between care values and the technical capabilities of a care robot – mirroring the approach of value-sensitive design. To deepen this understanding I use the approach of Akrich (1992) for highlighting the meaning attributed to the application of a care robot. The result is an understanding of the complexity of care practices and the recognition that design of care robots ought to proceed on a design-by-design basis given the impact that one capability vs. another has on the overall process of care.

Keywords care robots, care ethics, value-sensitive design, ethics and technology

Introduction

The latest in robotics development is the creation of robots for use in the care of elderly persons. I refer to these robots as care robots. A care robot may have any number or range of capabilities and may be used by the care-giver directly or by the care-receiver directly. For instance, a robot used by the care-giver in the practice of lifting a patient is a care robot. Alternatively, a robot used by a care-receiver as a reminder to take medications is also a care robot.

Given the ethically sensitive nature of the context and practices into which these robots will be stepping, ethical reflection is required. Attention is needed not only retrospectively, in terms of the impact of these robots, but also prospectively, in terms of how the design of these robots ought to proceed. In this

paper, I present a framework to be used for both the retrospective evaluation of care robot designs as well as inclusion in the design process of future care robots. The framework uses the blueprint of value sensitive design but differs in many respects. Namely, that its use is intended for the design of *any* care robot and not one in particular. Although the framework outlines the values of ethical significance to be taken into consideration in the design of a care robot, what is also required (on the part of the robots designer) is an understanding of the complex context within which care is taking place and how a difference of one capability vs. another can completely alter the promotion of values in a care practice. As roboticist Peter Asaro rightly points out, "*if one wishes to design a system to perform a task, then it is best to first understand that task*" (Asaro, 2009, p. 23). Accordingly, the importance of understanding the complexity of care tasks is necessary for understanding how a robot might fulfill said task, and for understanding how the technical capabilities of a care robot may threaten the promotion of values manifest through said task.

The aim of this paper is to present the main challenge for care robot designers: understanding the complexity of a care practice and its relationship with technical content. By 'understanding the complexity of a care practice', I mean understanding that values are abstract ideals and it is through the interactions and actions of actors that values come into being. Moreover, care is not just one value but is rather a cluster of values that come into being through actions and interactions. By 'understanding the relationship of this complexity with technical content', my aim is to show how the difference between one capability in a care robot can change the resulting care practice. In the following paper, I begin by articulating what a care robot is, what the proposed framework for the evaluation of a care robot is and how it is created, and why issues of design are so important at this time. Following this, I provide a conceptual investigation of values and of care to illustrate how I arrive at the criteria for the framework and in so doing I demonstrate their complexity. Finally, using examples of current care robots I show the relationship between care values and the technical content of a care robot. The result is an understanding of the complexity of care practices and the recognition that design of care robots ought to proceed on a design-by-design basis given the impact that one capability vs. another has on the overall process of care.

What is a care robot?

As I have mentioned, a care robot is one that is used in the care of persons in general. The definition of a care robot relies on the idea of *interpretive flexibility* (Howcroft, Mitev and Wilson, 2004), that a robot is defined by its context, users and task for use. This means that the same robot might be called by a different name if the robot is used for rehabilitation or for care purposes. The Hybrid Assistive Limb (HAL) is an example of this phenomena; the robot may be used in rehabilitation when worn by the patient or may be used to relieve the stress of lifting on the nurse. As such, the same robot is referred to as a rehabilitation robot or a care robot respectively. The demographic of focus for this work is elderly persons in the hospital, nursing home or home domain. These robots can be used directly by the care-giver in the care of another or may also be used by the care-receiver directly. There is no capability exclusive to all care robots – they may have any number and range of capabilities from planar locomotion (vs. stationary) to voice recognition, facial or emotion recognition. They may have any degree of autonomy, from human-operated to fully autonomous.

Creating a framework for the ethical evaluation of robots

Referring to the question of how to include ethics in the design process of care robots, the approach of Value-Sensitive Design (VSD) offers valuable insights (Friedman et al, 2006; Friedman et al, 2003). VSD has been praised for its success in incorporating ethics in the overall design process of computer systems or ICT (Van den hoven, 2007) but is also advantageous to guide the design process of a wide array of technologies (Cummings, 2002). In fact, in a recent paper of roboticists Sharkey and Sharkey, they recommend the approach of VSD as a method for the ethical design of robots (Sharkey and Sharkey, 2010). VSD takes as its starting point the belief that technologies, through their use, can actively promote certain societal values. In other words: a value is embedded in a technology such that through the use of said technology the value is promoted or demoted. For example, one might suggest that a bank machine promotes a value of autonomy or independence of citizens. While this technology might promote these values, at the same time it might inhibit others, namely the value of social interaction (a client no longer visits the human cashier). Thus, any technology may promote and inhibit different values at the same time.

Accordingly, technologies may be designed in a way that accounts for values of ethical importance in a systematic way and rigorously works to promote said values through the architecture and/or capabilities of a technology. It follows then that care robots may be designed in a way that promotes the fundamental values in care. Although VSD is meant for the design of a particular system or product, that is not my overall aim. My goal is to create a general framework that may be used by designers and/or ethicists in the ethical evaluation of a care robot, or for the inclusion of ethics in the design of a care robot. By using the blueprint of VSD I am creating a more specific framework that addresses the relationship between technical capabilities and design with the specific context of use, task of use and users in mind. An additional benefit to the framework is its potential use retrospectively and prospectively. When used retrospectively, designers are able to understand the impact of their design on the resulting care practice. When used prospectively, designers are able to incorporate the framework into the design process of a care robot, ultimately incorporating ethics into the design process.

The framework I am creating uses components of the VSD methodology in its creation – namely the conceptual investigation coupled with a brief empirical and technical investigation. As in traditional VSD, my conceptual investigation is an exploration of the value constructs of ethical importance. I diverge from traditional VSD in that I have selected the values which serve as the foundation of the healthcare tradition in Western cultures according to the World Health Organization. I attempt to understand how these values are interpreted philosophically by care ethicists, as well as how these values are interpreted in context through observational work. Given that the framework revolves around the relationship between values and technical content, I also incorporate a technical investigation by exploring technical capabilities of care robots currently available or available in the near future. Unlike the traditional empirical and technical components of VSD, I do not embark on empirical studies to test a care robot in context with human users (at least not at this moment in time). This is because I aim to provide a framework for the design of a *range* of care robots vs. one particular system. In order to utilise the framework it is necessary to shift back and forth from conceptual

to empirical to technical aspects, in much the same manner as VSD methods.

In Peter Asaro's article *What should we want from a robot ethic?* (2006), he proposes the three dimensions one could be referring to when one speaks of 'ethics of robots':

1. the ethical systems built into robots;
2. the ethical systems of people who design robots, and
3. the ethics of how people treat robots.

He then concludes that given the nature of robots as socio-technical systems, a framework for ethically addressing robots ought to include all three dimensions. For Asaro, the overarching question is that each of the three dimensions has to do with the distribution of moral responsibility in the social-technical network into which robots are introduced.

While Asaro presents a compelling case for the need for a comprehensive approach to robot ethics, he stops short of presenting such an approach. The framework I have created is intended to take up this challenge. I argue that this framework adequately addresses the three dimensions identified by Asaro as well as his overarching question. Asaro himself, however, seems to reject the approach of VSD. In a later paper written by him, *Modeling the Moral User* (2009), he claims that while VSD is useful as an educational tool, robots should not be evaluated on a design-by design basis, as the VSD approach would do, but rather the design of robots ought to be standardised based on the decision making capabilities the robot posses. In this paper I aim to show how the complexity of care tasks demands that the design of care robots be conducted on a design-by-design basis and hence that Asaro's aim of a standardised robot design is not feasible.

Why design?

The answer to the question of why one ought to pay so much attention to issues of design is grounded in three rationales. Firstly, there are no 'regulatory frameworks' for the design of robots outside the factory. These frameworks "*consist of legislation and technical standards, and interpretations thereof by certifying organizations. Operationalization of ethical criteria are given in these regulative frameworks for safety and sustainability criteria*" (Van Gorp and Van de Poel, 2008, p. 77). These frameworks are socially sanctioned and legitimised and thus provide the grounds for "*morally warranted trust in engineers*" (Van Gorp and Van de Poel, 2008, p. 88). Although such frameworks exist for technology design in general, there are none for the design of robots outside the factory. Consequently, the public is left without basis for trust in the design process as well as the resulting robots. Given the context within which these robots will be placed, trust in the engineers, design process and resulting technology is of the utmost importance and is a cornerstone of the healthcare tradition.

The second rational is closely aligned with the first. The technology is in its nascent stage of development and offers a unique opportunity for incorporating ethics further upstream in its development and throughout the design process. A variety of mandates and institutions exist in which interdisciplinary collaboration is deemed attractive for its capacity to stimulate discussion and reflection of the wider social and ethical questions pertaining to a technology. The CoTeSys (Cognition for Technical Systems) group at the Technical University of Munich or the Autonomous Systems and Biomechantronics lab at The University

of Toronto are examples of institutions bringing together computer scientists, engineers and psychologists for the development of robots. What's more, the field of robotics presents a unique opportunity to include questions of an ethical nature further upstream in its development in order to tailor both its development and implementation with an enhanced sensitivity to future outcomes – outcomes in terms of the broader social and ethical criteria. Interdisciplinary cooperation of this kind is what the Netherlands refers to as 'responsible innovation'. Given the context within which these robots will be implemented, the tasks for which the robots will be used, and the nature of the interaction between humans and robots, addressing questions of an ethical nature throughout the design process in an interdisciplinary manner is truly the responsible way to proceed.

Thirdly, from the perspective of philosophy of technology, many theories exist which seek to explain the reciprocal and dynamic relationship between society and the development of technologies. This may not be an explicit aim of the designer but is a condition of the work that they do. The theory of scripts illustrates how engineers' assumptions about user preferences and competencies show themselves in the technical content of an object (Akrich, 1992). Latour builds on this idea to show how technologies steer behaviours, moral and otherwise (1992). Verbeek shows how technologies are included in our decision making such that moral decisions are in fact a hybrid affair between humans and technologies (2006). In the computer ethics domain, Nissenbaum illustrates how values and biases are embedded into a computer system (1993). The golden thread through all of these perspectives is that social norms, values and morals find their way into technologies both implicitly and explicitly and act to reinforce beliefs or to alter beliefs and practices. Beyond the embedding of values and/or norms, once the robot enters a network it will alter the distribution of responsibilities and roles within the network as well as the manner in which the practice takes place. This shift is what Verbeek refers to as mediation: "*when technologies are used, they help to shape the context in which they fulfill their function, they help to shape human actions and perceptions, and create new practices and ways of living*" (Verbeek, 2008, p. 92). Akrich discusses this in terms of the assumptions designers have of the traditional and ideal distribution of roles and responsibilities – that practices may shift based on an assumption made by an engineer of how the practice 'ought' to take place, how roles and responsibilities 'ought' to take place and inscribing these assumptions into the technical content. It is these ideas that mirror the overarching question presented by Asaro – that a robot ethic ought to address the shift in responsibilities once the robot has been included into a socio-technical network. What's more, when a shift in roles and/or responsibilities is inscribed in a robot a valuation is being made – for example, that the human is not competent to fulfill the task or that the robot may fulfill the task in a superior manner. Thus, even assumptions about users may be considered statements of value at times.

It is true that the rationales presented here relate to the design of any system or technology; however, greater weight is added when one takes into account the context in which the care robot will be placed and the nature of the activities the care robot will fulfill. Without standards guiding the development of care robots how is one be sure that the values and norms central to the healthcare tradition will be promoted? Or, without making these norms and values explicit through the design process, how can one be sure their inclusion will be taken into account? Or, given the cost of development of these robots, mustn't they provide

the same quality of care as today if not better (which presupposes an understanding of how one defines 'good care')? Or simply, given the dramatic impact care robots may have on society, shouldn't future considerations be taken into account in design? With these rationales in mind, the design of any technology is ultimately a moral endeavour. The design of a care robot is even more so given the vulnerability of this demographic, the delicacy of their care needs and the complexity of care tasks.

Investigating the concepts of value and care

In order to understand the complexity of care in context and its relationship to the technical capabilities of a care robot, we must first explore the meaning of the terms value and care. With this understanding, we may then unearth the fundamental values of a care scenario which will ultimately allow us to expose the moral precepts to operationalise in the design of a care robot.

Values

According to the Oxford English Dictionary, values are conceived of as "*the principles or standards of a person or society, the personal or societal judgment of what is valuable and important in life*" (Simpson and Weiner, 1989). It follows then that when something is de-valued it loses importance. Values may be intrinsic/inherent to an object, activity or concept, or, things may be valued as a means to an end (Rosati, 2009). For example, in the healthcare context, the concept of human dignity is valued on its own whereas the activity of touch in care contexts is valued as a means to preserving the dignity of persons (Gadow, 1985). Things of value[155] may be valued on a personal level or on a societal/cultural level. Different cultures or groups have different meanings and interpretations of values as well as what counts as being valuable. Values may also be more of a subjective enterprise (various things valued for an individual) or more of an objective enterprise (universal values such as justice, human dignity, fairness) although the premise that universal values exist may also be contested. The latter does not imply that values considered abstract and universal are interpreted in the same way between cultures, groups, or time periods but rather that the valuation of things may differ from an individual's sphere to a more public one. Linked with the concept of 'good' a value may be construed as something that is good or brings about a good consequence. In this way, a value is a rational construction that helps guide one in their moral decision making or judgments.

In the VSD literature, Batya Friedman and colleagues, opt for a more open definition of a value to refer to "*what a person or group of people consider important in life*" (Friedman, 2003, p. 2). This implies then that

[155] I have used the word 'things' here to bypass repeating people, places, activities, concepts, and objects, all of which are included in the discussion of values.

all values are not interpreted in the same way. Nathan et al illustrate this with the value of privacy and its divergent ways of being interpreted and manifest between cultures (2008). Le Dantec et al reinforce the idea that values may be universal, or generally accepted, but differ in their interpretation. Because of this, Le Dantec et al suggest a way in which the methodology of VSD may be strengthened, through an uncovering of values *in situ*, or discovering values through experiencing the practice (Le Dantec et al, 2009). This is of course due to the idea that differences exist between designers' values and users' values (Nathan, 2008). Thus, the scope of values varies depending on the technology, the users, the culture, the time period and the application domain. In the VSD methodology, Friedman selects the values of ethical importance pertaining to computer systems. Given that my framework is intended for use in the design of care robots, the values pertaining to the specific context are of greater ethical significance and relevance.

Defining care and care values

Care may be one of the most difficult concepts to articulate. This is in part due to the ubiquity of the word but is also largely a consequence of the fact that one is assumed to know what care means given its revered place in many cultures. The work of Warren T. Reich nicely outlines the broad range of meanings and connotations care has embodied going back as early as Ancient Greece (Reich, 1995). Regardless of how one perceives or defines care, care is valued as something above and beyond simple care giving tasks. It has a central role in the history of humankind as a means to signify the value of others. In other words, by caring you bestow value on the care-receiver.

In the verb 'to care' one finds that caring may actually be divided into the idea of *caring about* and *caring for*. The dimension of *caring about* in the medical field implies a mental capacity or a subjective state of concern. On the other hand, *caring for* implies an activity for safeguarding the interests of the patient. In other words, it is a distinction between an attitude, feeling or state of mind vs. the exercise of a skill with or without a particular attitude or feeling toward the object upon which this skill is exercised (Jecker, 2002).

Alternative to the idea that care in itself is a value – linked with the good life and with a valuation of another – care is in fact a cluster of many other values. These values are given importance for their role in care – their role in giving significance to care, in making care what it is. These values form the buttress for care as an ethical endeavour and create a framework for evaluating care as a practice. It is through the manifestation of these values that one comes to understand what care really is in practice. It is therefore fruitful for the topic of embedding care values to understand these values and their link with consequences. Thus, to begin from a top-down approach, I look to the values articulated by the governing body of healthcare, namely the World Health Organization (WHO). The WHO framework for people-centered health narrows in on the values in healthcare stemming from the patient's perspective: *patient safety*, *patient satisfaction*, *responsiveness to care*, *human dignity*, *physical wellbeing* and *psychological wellbeing* (2007). This is not to say that other values like innovation or physician autonomy are not respected but rather from the patient's perspective the listed values are the ones with the greatest ethical importance and will thus be used in my evaluation of implementing robots in the care of persons.

Given the abstract nature of values presented in this way, I take the suggestions of Le Dantec et al

(2009) to understand the specific interpretation of values in context. To do this I completed fieldwork experience in a nursing home[156]. Interestingly, the interpretation of values as well as their ranking and meaning differed depending on:

1. the type of care (eg. social vs. physical care);
2. the task (eg. bathing vs. lifting vs. socializing);
3. the care-giver and their style, as well as the care-receiver and their specific needs.

In support of the values identified through the WHO, the mission statement of the nursing home includes additional values such as *compassion*, *integrity*, *dedication*, *respect* and *accountability*. We can see then that the more abstract values articulated by the WHO are made specific when put into context. This means that, for care tailored to the needs of elderly persons, accountability is a value to be upheld and most likely is a manifestation of the WHO values of patient safety, patient satisfaction and perhaps human dignity. Compassion is also highly valued in the context for this demographic. Although compassion is not stated by the WHO, one might assume that compassion is also a manifestation of a WHO value, again human dignity. Compassion is an attribute of the staff providing the care that reflects the concept and value of *caring about.* The values of this mission statement presume that personalised care is a value and priority – this type of care requires respect on the part of the care-giver for the integrity of the care-receiver's individual spiritual and cultural beliefs. Although personalised care is not an explicit value of the WHO, it may be considered a manifestation of any of the WHO values.

Of great significance is that all of the values central to the healthcare tradition are observable within the relationship between the nurse and client. Not only is the relationship the place where values are expressed and promoted, but there are also certain components of the relationship which, when manifest, are valued as integral mechanisms for good care. These components are: power, trust, respect and intimacy[157]. These components may not all be considered values so to speak, but are *valued* given that their expression symbolises a manifestation of another value. For example, power is not a value in the same way as safety or client choice, but sensitivity to the unequal power within the care-giver + care-receiver relationship – the power of the nurse and the vulnerability of the client – represents a valuation of the other's integrity and/or dignity. As such, attentiveness to these components is valued. It follows that the relationship is valued on its own but also as a way of manifesting many of the other values central to health care and to care in general.

---

[156] Fieldwork experience was gained by volunteering in a nursing home in London, Ontario, Canada for 4 weeks as a 'life enrichment coach'.

[157] Taken from *Standards for the Therapeutic Nurse-Client Relationship*; the Royal College of Nurses of Ontario, revised 2006, http://www.cno.org/Global/docs/prac/41033_Therapeutic.pdf?epslanguage=en

What's more, touch is an important action in care that is valued on its own as well as a means for manifesting other values like respect, trust and intimacy. Touch is the symbol of vulnerability, which invokes bonds and subjectivity (Gadow, 1985). Touch acts to mitigate the temptation for objectification. Thus, touch is considered a value in the healthcare domain, the outcome of which results in the preservation of the value of human dignity. Using the value of touch as an example, we can see then how a certain technology might impede its manifestation. Melanie Wilson illustrated how a particular computer system implemented in the field of nursing was rejected as it prevented nurses from 'hands on care' – from touch – a cornerstone of the nursing tradition (Wilson, 2002).

Not only is there a broad range of values integral to the healthcare tradition, but a broad range of interpretations and prioritisations of these values. How a value is interpreted is dependent on both the context and the personal experience of the care-receiver and care-giver. For example, in a ward with people suffering from dementia, safety means not letting patients wander onto the streets, or preventing patients from hurting both themselves and others. In a 'typical' ward of a nursing home, safety is in terms of preventing patients from falling, or assisting in the feeding of patients to prevent chocking. In addition to context, how a value is interpreted and prioritised is also dependent on personal experiences as well as the specific practice. For example, through the practice of lifting, the value of safety is manifest (or interpreted) through ensuring the care-receiver does not fall or is not injured. Here, safety is of paramount importance. In contrast, through the practice of bathing, the value of safety is interpreted in terms of suitable water temperature (not burning or scarring the patient), and proper positioning on the bed or in the tub to prevent injury. In the practice of bathing, however, while safety is of the utmost importance, other values take precedence. For example, closing the curtain to ensure privacy, verbal communication to calm the care-receiver, and gentle strokes to convey empathy and respect through the practice.

In short, not only is care a value for what it symbolises (a valuation of another) and manifests (meeting the needs of another) but it is also valued for the elements that make up care; elements like trust, respect, intimacy and respect for the asymmetry in power. This investigation was meant to shed light on the complexity of care, care values and how these values are interpreted. What was also evident was the intertwining of care values – the expression of one often presupposes the expression of another. To expand on the intertwining of values, care ethicists often speak in terms of care practices.

Care practices

To elaborate on the marriage between *caring about* and *caring for*, a useful concept is that of a care practice. A care practice is, as care ethicist Joan Tronto describes it, a way to envision a care task or a series of care tasks [please insert a reference!]. A way in which one can grasp the fortitude of each action and interaction between a care-giver and care-receiver. More importantly, it is a way to envision the holistic nature of care.

> The notion of a care practice is complex; it is an alternative to conceiving of care as a principle or as an emotion. To call care a practice implies that it involves both thought and action, that thought and action are

interrelated, and that they are directed toward some end. (Tronto, 1993, p. 108)

Understanding that care tasks are more than just 'tasks' but are rich practices in a value-laden milieu that act to bring about the promotion of values, may be one of the most crucial points for designers to grasp. The reason for this has to do with understanding how values are manifest and thus how a design will impact this materialisation of values. To exemplify this shift from task to practice, let me use the practice of lifting. When a patient is lifted by the care-giver, it is a moment in which the patient is at one of their most vulnerable. The patient trusts the care-giver and through this action a bond is formed and/or strengthened which reinforces the relationship between care-giver and care-receiver. The significance of this is apparent in the actual practice of lifting but comes into play later on in the care process as well. Trust, bonds, and the relationship, are integral components for ensuring that the care-receiver will comply with their treatment plan, will take their medication and be honest about their symptoms. Without trust, these needs of the care-giver are threatened, ultimately threatening the entire care process and the good care of the care-receiver. Thus, conceptualizing care tasks as practices adds a deeper meaning to each 'task'. It is within a care practice that the values are manifest and given their significance but it is also within practices that the holistic vision of care takes form.

While many care ethicists make clear the range of values and principles which provide a normative account for care, they fall short of providing a systematic way to *visualise* and *evaluate* these principles and values. The vision presented by Joan Tronto allows for a perception of care as a process with stages and elements which provides the most enticing conceptualisation for engineers to work with. There are four phases of a care practice for Tronto:

1. Caring about (recognizing one is in need and what those needs are);
2. Care taking (taking responsibility for the meeting of said needs);
3. Care giving (fulfilling an action to meet the needs of an individual);
4. Care receiving (recognition of a change in function of the individual in need).

These phases have corresponding moral elements as standards to evaluate the care practice from a moral standpoint. These elements are:

1. Attentiveness;
2. Responsibility;
3. Competence and
4. Responsiveness.

Attentiveness refers to an attribute or virtue of the care-giver, a certain competence for recognizing needs. *Responsibility* refers again to an element of the care-giver and their stance or concern for ensuring the care-receiver is pointed in the right direction for care or maintaining an accurate assessment of needs etc. Responsibility is often delegated to a moral agent; however, some responsibilities are delegated to an artefact, or are shared with an artefact, as technologies are wide spread in healthcare. Here, the concept of mediation (Verbeek, 2003) becomes critical in the sense that decision making on the part of nurses and patients is a hybrid affair between the nurse/patient and existing technologies. *Competence* is once again an attribute of the care-giver and refers to the skills with which the care is given. An unskilled care-giver may be

more detrimental than no care at all. *Responsiveness* refers to an attribute of the care-receiver and their role in the relationship – to guide the care-giver – but also refers to the openness of the care-giver to the impact of care on the care-receiver. As such, each of the participants addresses the other. This element (and the phase of care receiving) is important for remembering the reasons for care in the first place: the care-receiver and their need. Without this, care is not complete. This recognition also encourages an active stance of the care-receiver rather than a more passive, vulnerable one.

Care practices and care values

A care practice is the attitudes, actions and interactions between actors, human and nonhuman, in a care context that work together in a way that manifests care values. Human actors are the care-giver and care-receiver as well as a range of other healthcare professionals and perhaps the family or loved ones of the care-receiver. Nonhuman actors range from the mechanical bed, the sink, the window, the curtain enclosing the patient and so on. Thus, a care practice is defined by the interactions between actors; the practices are values working together. For example, the element of attentiveness is required for recognizing the concrete needs of a patient essential for meeting the WHO value of patient satisfaction. Attentiveness often demands the assistance of a technology (a nonhuman actor) in association with a care-giver for greater accuracy. The element of competence ensures that the values of physical and psychological wellbeing will be met with skill. The element of responsibility corresponds with the value of accountability for the nursing home.

The meaning of a practice is found not only in the actions of a care-giver but in *how* the actions and interactions take place. Care-giving through bathing, lifting, or feeding is considered good care when done with consideration for the care-receivers vulnerable state. This may be seen through the tone and level of volume the care-giver uses to speak or the gentleness with which the care-giver touches while bathing or lifting. A care-giver must be attentive to the particular preferences of one care-receiver and another and must be competent in executing care in a preferred manner. Thus, actions are not only valued on their own but are also dependent on the manner of their manifestation.

In short, each of the values articulated by the nursing home mission statement (compassion, integrity, dedication, respect and accountability) presume the elements of attentiveness, responsibility, competence and responsiveness. The element of attentiveness is required for recognizing; the vulnerability of the patient, the unequal distribution in power, when a patient is in need of touch and how much (meaning a slight pat on the back or a hand hold) or when a patient would prefer not to be touched. The element of responsibility is closely aligned with the value of trust in the care-giver + care-receiver relationship. What's more, trust is a necessary condition for the manifestation of values like touch. Consequently, we see not only the intertwining of values, actions and interactions but also that the elements are the forum within which the values are manifest. Some elements are analogous with values while others act as the vehicle for the promotion of a value. Either way, the elements encompass all of the aforementioned values.

Selecting the values of ethical importance in care

Essentially, care is a cluster of values that come into being through the actions and interactions of actors in a care context. Creating a standardised framework to guide the promotion of these values which applies to any care context, task, care-receiver or care-giver reveals itself to be quite problematic given the range and variety of care values discussed in the former section. In other words, to claim that human dignity, compassion or respect for power are values to be embedded in a care robot offers nothing for the designer in terms of the robot's capabilities. Moreover, their ranking and prioritisation is dependent on the context (e.g. one hospital domain or another vs. a nursing home) and task (e.g. lifting vs. bathing). However, in the care ethics literature, alongside values, *needs* play a central and crucial role in the provision of good care. The needs of the patient mark the starting point of the care process and the process then revolves around a care-giver taking steps to meet these needs. Understanding the multiple layers of needs, the many ways in which they might be fulfilled, the preferences for one way over another, and the divergent needs between individuals adds a further complexity to the meeting of needs. If this wasn't complicated enough, the care-giver has needs too! (S)he has needs in terms of resources, skills, responsiveness from the care-receiver to understand when needs have been met as well as their own personal needs.

Given the central role of needs in a care context, what might the relationship be between needs and values? Although many authors have written on the subject, little consensus can be found. I suggest then that values in healthcare are given their importance for their role in meeting needs. This corresponds with Super's conceptualisation of the relationship between needs and values: "*values are objectives that one seeks to attain to satisfy a need*" (1973, p. 189-190). This means that, the value is the goal one strives towards and in so doing, intentionally meets a need. In other words, we begin with needs, and the values represent the abstract ideals which, when manifest, account for the needs of individuals. It follows then that a framework for designing care robots ought to address the meeting of needs. Unfortunately, I've just shown how multifaceted and intricate needs are for the care-giver and care-receiver. What's more, according to the field of care ethics, it is neither possible nor advisable to outline a series of needs which pertain to all care-givers, care-receivers or care tasks in every instance/scenario (Tronto, 2010). While useful for policy, it goes against the vital element in care – that of the individual and their unique, dynamic needs. In other words, care is only thought of as good care when it is personalised (Tronto, 1993). There is, however, a solution to this barrier. It is possible to delineate a set of needs for *every care practice*. To recapitulate, together the phases and the moral elements make up a care practice. The practices are values working together and the vehicle for this lies in the moral elements. If we assume a care practice ought to proceed according to Tronto's phases, then the needs for every care practice are the corresponding moral elements. It is therefore these elements that ensure the promotion of care values. Consequently, it is these elements – attentiveness, responsibility, competence, responsiveness that make up a core portion of the framework.

With this suggestion, there are two assumptions being made:

1. That every care practice will ALWAYS have the moral elements as needs, independent of the care-giver and care-receiver, and
2. That the values are subsumed within the moral elements.

Using the practice of lifting as an example to illustrate the first assumption, I am making the claim that this practice will ALWAYS require attentiveness, responsibility and competence on the part of the care-giver and will ALWAYS require a reciprocal interaction between care-receiver and care-giver for determining whether or not the needs have been met, no matter what the context is. In this way we can see that the moral elements are needs which are independent of the care-giver and the care-receiver. They are, however, dependent on the context and the specific practice for their interpretation and prioritisation. If we were to compare the practice of lifting with the practice of feeding we would see how the element of competence is uniquely interpreted in each practice (skilfully bearing the weight of another without dropping or causing pain vs. skilfully coordinating timing and placement of food and utensils). In terms of context, the practice of lifting in the hospital requires greater efficiency than the practice of lifting in a home setting where time may not be as much of an issue. Thus, although the moral elements must always be present, context and practice still play a crucial role in their interpretation, prioritisation and manifestation. This recognition is also included the framework.

For the second assumption – that the values are subsumed within these moral elements – the values are often analogous to a phase or moral element or are expressed through the manner in which an action takes place. The value of patient safety is fulfilled through the competent completion of a practice (the phase being care giving and the moral element being competence). The valued action of touch requires attentiveness on the part of the care-giver for determining when and to what degree touch is considered necessary. The manner in which care practices take place is often tailored to the specific likes of one care-receiver or another and again requires attentiveness to those preferences and competence in meeting them. What's more, paying attention to those unique preferences is a vehicle for establishing trust and allowing for successful reciprocal interaction.

In short, ensuring the elements are present or strengthened through the design and introduction of a care robot ultimately results in a manifestation of the core care values. The differences in the prioritisation and manifestation of the elements between practices and/or contexts is something that the care ethicist may draw the attention of the designer to. But the designer must first be aware of the necessary elements and their manner of manifestation.

The framework

Given the central and crucial role the design of care robots will play on the resulting care practice and the overall provision of good care, addressing ethical issues throughout the design process is pertinent. As identified through the conceptual investigation, good care is first and foremost personalised. The standards for evaluating good care are dependent on: context, the individual care-receiver and care-giver, fulfillment of care values, and the meeting of needs through care practices. It follows that a framework for the ethical assessment of care robots ought to incorporate these dimensions.

Context – hospital (and ward) vs. nursing home vs. home etc

Practice – lifting vs. bathing vs. feeding vs. delivery of food and/or sheets…[158]

Actors involved – nurse and patient vs. patient alone vs. nurse…

Type of robot – assistive vs. enabling vs. replacement…

Manifestation of elements - Attentiveness, responsibility, competence, responsiveness

In the textbox I've outlined the criteria for inclusion in the framework. One must first identify the context within which the care practice is taking place (e.g. hospital, and the ward, vs. nursing home vs. home), the practice for which the care robot will be designed (e.g. lifting vs. bathing vs. feeding vs. delivery food or sheets to the room) and the actors involved (nurse, patient, mechanical bed, curtain etc). For example, a task like lifting traditionally involves the patient, nurse (or porter), mechanical bed, mechanical lift, and curtain enclosing the patient. With this in mind the next task is to conceptualise how the moral elements come into play for that particular practice in that particular context. In retrospective evaluations, the section 'manifestation of elements' refers to the impact the robot has on the traditional manifestation of elements. For example, a human enabling robot (that acts as an aid to the nurse) will have a different impact on the moral elements than a replacement robot (one that replaces the nurse as an actor). With respect to the prospective design (i.e. when the framework is included in the design process of the care robot), 'manifestation of elements' refers to how the robot acts to promote these elements or alternatively how the robot might impede the promotion of these elements. This may be similar to how a human traditionally does so or may be thought in terms of how the elements differ when translated into technical capabilities. Therefore, the case-by-case design process that I made reference to at the beginning of this paper is not user-by-user or context-by-context but practice-by-practice while incorporating both context and users.

[158] The ellipsis (sequence of dots) following the description of a criterion indicates that the list is not exhaustive and may include additions.

Using the framework for the retrospective evaluation of current care robot designs

As I have claimed, the framework may be used for both the retrospective and the prospective ethical assessment of care robots. In order to demonstrate how it may be used for the retrospective ethical evaluation of care robots – specifically how the capabilities of a care robot impact the practice of care – I take the practice of lifting in a nursing home context and compare two care robots which may be used for this practice.

The practice of lifting in the nursing home

One of the more challenging practices for the nurse is the lifting of patients. Many elderly patients in the hospital or nursing home require partial assistance for lifting themselves out of bed or out of a chair. Alternatively, many are not capable of supporting their own weight at all and require complete assistance of a nurse to get out of bed or out of a chair. Given that the nurse must do this for any number of patients, there is a risk to the nurse's physical safety if she/he is required to lift every patient. What's more, many nurses are not physically strong enough for this. As a result, nurses have opted to use mechanical lifts on the many occasions that patients need to be lifted (Li J et al, 2004).

The first wave of automation presented a rather flat view of the care practice of lifting. It appears to have viewed the practice as a task, as an event that is separate from the process of care and uninvolved in the manifestation of care values. For example, using the mechanical lift for complete assistance, the patient is lifted using a remote control, controlled by the nurse. The patient is then lowered into the chair. When the patient is being lifted there is no physical contact with the nurse, although the nurse is physically present there is no chance for eye contact as the patient is raised quite high and the nurse is paying attention to the remote control. Thus, eye contact and touch are not possible. As I have already shown, these values are integral for establishing and/or maintaining a trusting bond and this bond is integral for the provision of good care later on in the process (the patient complying with their treatment plan, taking medications, being honest about their symptoms etc).

The current technology involved in the practice of lifting shows us how important it is for designers to understand the holistic vision of a care practice – how it acts as a moment for the promotion of care values. Consequently, the introduction of care robots presents a unique opportunity to re-introduce certain values of ethical importance. Alternatively, a robot may perpetuate the trend to minimise certain care values.

Care robots for the practice of lifting

There are two robots which will be used to show the utility of this framework in the retrospective ethical evaluation of current care robot designs. The first is the RI-MAN robot from the Riken Institute (Onishi, 2007) and the second is the Hybrid Assistive Limb (HAL) from Cyberdyne (Hayashi, 2005). Both robots can achieve the same task (lifting a patient); however the technical capabilities through which this task is achieved differ and thus change the way in which the caring task is fulfilled along with the resulting care

scenario. The RI-MAN robot is an autonomous robot, meaning it is capable of lifting a patient and carrying him/her from one place to another without being controlled by a human operator. This robot is designed to work directly with humans and as such is programmed for safety considerations like speed as well as the materials which are used for its structure. The robot has a humanoid appearance, meaning the robot has a head, eyes, a nose and arms. This robot can work in a hospital, a nursing home or in someone's home.

The second robot, HAL, is an exoskeleton, meaning a human operator wears the robot in order for it to fulfill its task. The robot is a weight displacing robot such that the human does not feel the full effects of the weight. Versions of this type of robot exist in factory and military applications to prevent over-exertion of factory workers or soldiers respectively. It is not an autonomous robot, but a human-operated one. It too will interact directly with a human (more than one in most instances) and must be programmed for the appropriate safety considerations. Given that the robot is human-operated, the safety considerations for this robot are slightly different from those of the RI-MAN. For example, the robot will not have the same sensors for perceiving a wall, person or object in its range. This robot, in contrast with the first, does not have a humanoid appearance, but appears rather machine like. This robot can also be used in the hospital, nursing home or home. While the previous robot is capable of replacing the human care-giver that would normally lift the patient, this robot is meant to assist the human care-giver with their task. By reading the biometric signals of the care-giver, the robot is able to bear the burden of the weight of whatever the care-giver is lifting. This could be a patient, a bed, a heavy box etc. We can see with this robot that if used for the rehabilitation of a patient unable to walk it is a rehabilitative robot (Kawamoto, 2002), whereas, if it is used in the hospital it is considered a type of care robot.

Reflections of design on the elements in care

When comparing the two robots and their impact on the elements in care, we might suggest that in the case of the RI-MAN (autonomous) robot, all elements have been delegated to the robot. This means that the robot is responsible for being attentive to the frailty of the patient when lifting; the robot is ultimately responsible for safety throughout the practice; the robot is required to fulfill the practice in a skilful manner, and the robot is responsible for perceiving whether the needs of the patient have been met. A consequence of the application of this robot is a threat to the holistic process of care. Given that the robot replaces the care-giver for this practice, it may threaten the element of trust which is required further along in the process of care. This is not to say that trust cannot be established through another practice, but rather that it does not present the forum in which trust is normally established or strengthened. In terms of values like compassion, respect, and integrity, promoted through the human care-giver, it is possible to suggest that these values come into being exclusively through an interaction between humans. Touch, as an example of an action that helps to establish and promote values like compassion, respect and integrity, is missing. Eliminating human-to-human touch suggests that the design of this robot perpetuates a de-valuing of touch making the application of this robot questionable. Alternatively, there are care-receivers in their homes who would prefer the assistance of an impartial robot over a spouse to keep their dignity and integrity intact. Consequently, having a robot to fulfill the practice of lifting may be seen as a more compassionate means when the care-

receiver's vulnerability is maximised by requiring help for these practices. This divergence shows the importance of context in the ethical assessment of a care robot.

In the case of HAL, the element of attentiveness is still in the domain of the human as is the element of responsiveness. For the former, the care-giver uses their own faculties to ascertain when the care-receiver needs to be lifted, at what speed, from which angle and with or without social interaction. For the latter, responsiveness or reciprocity is something that happens between the care-giver and care-receiver in real time by verbal and nonverbal cues detected by the care-giver. This means that the nurse can ask the patient how they are doing while they are lifting. As for responsibility and competence, these elements now become shared endeavours between the human and the robot given that the weight is being displaced to the robot. The care-receiver and care-giver must both trust the technology – responsibility for the safety of the practice becomes a hybrid event between the human and robot. Additionally, a certain amount of competence for the skilful completion of the practice is delegated to the robot. Thus, a portion of the responsibility for lifting is delegated to the robot as is a certain level of skill. But this is done in an assistive way, thus the human care-giver is still responsible overall. With RI-MAN the role and responsibility of trusted care-giver is entirely delegated to the robot whereas with HAL, the role and responsibility of care-giver is a shared effort between human and technology.

Attributing meaning to design through assumptions

It is only through a deeper understanding of what care values are and how they are manifest throughout a care practice that we come to grasp the impact a design might have on the care practice. Above and beyond the direct relationship one might uncover between care values and the technical capabilities of the care robot, there is greater meaning attributed to these capabilities upon further reflection. This meaning may only be grasped when one understands the intricate details of the care practice.

Akrich discusses the embedding of elements in terms of assumptions made about user preferences and competencies (1992). Placed in context, each robot takes on a distinctive meaning and the meaning of the robot has to do with the assumptions embedded within. This description is quite useful for my reflection and an important distinction must be made here pertaining to the difference between assumptions and the concept of values and norms. *Assumptions* are more about the real word, they are descriptive in a sense while *values* are more about what the real world ought to be like, they are normative in a sense. When an assumption is made about a value to be embedded, it does not have to be a description about what is, but could also be a claim about what values ought to be expressed, how they ought to be expressed, or what priority they ought to be given. In others words, when the built-in assumption pertains to a value, or when a valuation is being made, the result is a normative claim about what the values should be, what should be valued, or what the ideal is. For Akrich, "*many of the choices made by designers can be seen as decisions about what should be delegated to a machine and what should be left to the initiative of human actors*" (1992, p. 216). By making choices about what should and should not be delegated to certain actors (human or nonhuman), engineers may change the distribution of responsibilities in a network. Or, as Verbeek claims, engineers are materializing morality (2006).

Consequently, each robot reflects a divergent vision of the understanding of a care practice, the aim of the care practice and the prioritisation of values manifest through a care practice. RI-MAN represents a vision of the practice of lifting not requiring any of the values traditionally involved; touch, eye contact, human presence. One might suggest that the practice of lifting without any of these criteria is viewed in a rather flat way, as a task. Here, the ideal care practice is a standardised one where the value of efficiency is placed as the top priority. Although efficiency was not explicitly discussed previously, it is thought to fall under the realm of competence. This one-dimensional view of good care as efficient may have negative implications for the overall care process. One may presume that the quality of interactions, the number of social interactions, and the presence of a human are threatened by this efficient system. Alternatively, the system may be considered efficient considering that time of the human care-giver is freed up, ultimately improving the number of social interactions and the quality thereof. Thus, both design and integration into the healthcare system are of importance here.

Alternatively, HAL reinforces the elements of the framework as being integral needs of the care practice of lifting. This robot pays tribute to the holistic vision of care and the intertwining of needs and values as being expressed through a variety of practices. The vision of care presupposed in the design of the HAL care robot is one in which individualised care with a human care-giver present at all times for all parts of the care practice, is the overall aim. Efficiency is still a priority; however, it is achieved through meeting the need of a care-giver by contributing to the element of competence (enhancing the skill with which the care-giver may perform their role), attentiveness, (enabling the care-giver to perceive the minute cues of the care-receiver through the practice of lifting), and responsiveness (closely aligned with attentiveness but also embodies the reciprocal dimension of the relationship).

I cannot say whether this is the epistemic aim of engineers, but can only point to the potential meaning that the robot may take on through pervasive use, and the presupposing assumptions directing such a meaning. Moreover, this is not to say that RI-MAN ought to be disregarded or labelled as unethical. A different context might change things. For instance, in the home of two elderly persons who may not be equipped for wearing HAL or who may not want to burden their spouse when it comes to lifting, RI-MAN may be the more suitable, ethical choice. Clearly, decisions concerning the use of a robot and its ethical implications are many-sided and complicated and demand an understanding of the specific context and users for anticipating how the elements will be served to their greatest potential.

Conclusion

The prospective robots in healthcare intended to be included within the conclave of the nurse-patient relationship require rigorous ethical reflection to ensure their design and introduction do not impede the promotion of values and the dignity of patients at such a vulnerable and sensitive time in their lives. The ethical evaluation of care robots requires insight into the values at stake in the healthcare tradition. Only by understanding the complexity of care are designers able to uncover the deeper meaning the robot may take on when put in context.

Given the stage of their development and the lack of standards to guide their development, ethics

ought to be included into the design process of such robots. The framework I have suggested here takes the elements of moral relevance as its structure – as a way to systematically account for the values in a care practice. These elements, which may be considered needs of any care practice, are the manner in which values are manifest through actions and interactions of the care-giver and care-receiver. The framework is general in that it cannot standardise the creation of care robots. This is not possible for a variety of reasons. The first being that the standardisation of care goes against the most fundamental principle and core value of care – individualised/particularised care to an individual's dynamic needs in a specific context. Second, the capabilities of the robot will differ depending on the practice for which the robot is intended. Third, the capabilities of the robot and the robot's control will also differ depending on who the robot is intended to be used by (the care-giver, the care-receiver or a combination of the two). Fourth, the capabilities of the robot will differ depending on the goal of said robot. This means that the robot may be used as a support, reinforcement, enabler or replacement for a certain capability.

This paper attempted to show why a focus on design is the way to pursue the development of care robots and how the complexity of care tasks demands that robots be designed on a design-by-design basis. To illustrate the utility of the framework from an ethical perspective as well as a design perspective, I used two current designs of care robots to show how their different designs reflect divergent visions of care and the values inherent to care. By comparing the two robots – on a design-by-design basis – we may then decide which kind of care we want to provide. Consequently, the proposed framework allows us to systematically structure the design and development of care robots and as such provides the most promising approach for the ethical design and use of future care robots.

References

Akrich, M. (1992). The de-scription of technical objects. In W. Bijker, and J. Law (Eds.), *Shaping technology/building society*. Cambridge, MA: MIT Press.

Asaro, P. (2009). Modeling the moral user: Designing ethical interfaces for tele-operation. *IEEE Technology & Society*, *28*(1), 20-24.

Asaro, P. (2006). What should we want from a robot ethic?. In R. Capurro and M. Nagenborg (Eds.), *Ethics and robotics*. Amsterdam: IOS Press.

Brey, P. (2009). Values in technology and disclosive computer ethics. In L. Floridi (Ed.), *The Cambridge handbook of information and computer ethics*. Cambridge: Cambridge University Press.

Cummings, M. (2006). Integrating ethics through value sensitive design. *Science and Engineering Ethics 12*(4): 701-715.

Engelberger, J. (1989). *Robotics in service*. MIT Press.

Friedman, B., Kahn, P. Jr. et al. (2006) Value sensitive design and information systems. In P. Zhang and D.

Galletta (Eds.), *Human-Computer interaction in management information systems: Foundations* (348-372). New York: M.E. Sharpe.

Friedman, B. and Kahn, P. Jr. (2003) Human values, ethics, and design. In E. Sears and J.A. Jacko (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications* (1177-1201). Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc.

Gadow, S.A. (1985). Nurse and patient: The caring relationship. In A.H. Bishop and J.R. Scudder Jr. (Eds.), *Caring, curing, coping; Nurse, physician, patient relationships* (31-43). Alabama: The University of Alabama Press.

Hayashi, T., Kawamoto, H., & Sankai, Y. (2005). Control method of robot suit HAL working as operator's muscle using biological and dynamical information. In *IEEE International Conference on Intelligent robots and systems* (3063-3068).

Howcroft, D., Mitev, N. & Wilson, M. (2004). What we may learn from the social shaping of technology approach. In J. Mingers and L.P. Willcocks (Eds.), *Social Theory and Philosophy for Information Systems* (329-371). Chichester: Wiley.

Jecker, N.S., Carrese, J.A., & Pearlman, R.A. (2002). Separating care and cure: An analysis of historical and contemporary images of nursing and medicine. In E. Boetzkes & W.J. Waluchow (Eds.), *Readings in healthcare ethics* (57-68). Canada: Broadview Press.

Kawamoto, H., & Sankai, Y. (2002). Power assist system HAL-3 for gait disorder person. In *Computers helping people with special needs; Lecture notes in computer science 2398,* 19-19.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. Bijker and J. Law (Eds.), *Shaping technology/building society*. Cambridge, MA: MIT Press.

Le Dantec, C., Poole, E., & Wyche, S. (2009). Values as lived experience: Evolving value sensitive design in support of value discovery. In *Proceedings of the 27th International conference on human factors in computing systems.*

Li, J., Wolf L., & Evanoff, B. (2004). Use of mechanical patient lifts decreased musculoskeletal symptoms and injuries among health care workers. *Injury Prevention 10*(4): 212-216.

Lofquist, L., & Dawis, R. (1978). Values as second-order needs in the theory of work adjustment. *Journal of Vocational Behavior* 12: 12-19.

Nissenbaum, H. (1998). Values in the design of computer systems. *Computers and Society*: 38-39.

Onishi, M., ZhiWei, L., Odashima, T., Hirano, S., Tahara, K., Mukai, T. (2007). Generation of human care behaviours by human-interactive robot RI-MAN. In *IEEE International conference on Robotics and Automation* (3128-3129).

Programme for Responsible Innovation: Netherlands Organisation for Responsible Research. Retrieved from http://www.nwo.nl/nwohome.nsf/pages/NWOA_73HBPY_Eng.

Reich, W.T. (1995). History of the notion of care. In W.T. Reich (Ed.), *Encyclopedia of Bioethics.* Revised edition (319-331). New York: Simon & Schuster Macmillan.

Rosati, C. (2009). Relational good and the multiplicity problem. *Philosophical Issues 19*(1): 205-234.

Sharkey, A., & Sharkey, N. (2010). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics of Information Technology*. Retrieved from http://dx.doi.org/10.1007/s10676-010-9234-6.

Simpson, J.A., & Weiner, E.S.C. (1989). *The Oxford English Dictionary.* Oxford: Clarendon Press.

Super, D.E. (1973). The work values inventory. In D.G. Zytowski (Ed.), *Contemporary approaches to interest measurement.* Minneapolis: University of Minnesota Press.

Tronto, J. (2010). Creating caring institutions: Politics, plurality, and purpose. *Ethics and Social Welfare*, *4*(2): 158-171.

Tronto, J. (1993). Moral boundaries: A political argument for an ethic of care. New York/London: Routledge, Chapman and Hall Inc.

Van den Hoven, J. (2007). ICT and value sensitive design. *International Federation for Information Processing 233*: 67-72.

Van Gorp, A., & Van de Poel, I. (2008). Deciding on ethical issues in engineering. In . P. Vermaas, P. Kroes, A. Light, S. Moore (Eds.), *Philosophy and design: From engineering to architecture* (77-104). Springer.

Van Wynsberghe, A. (2011). Designing robots for care; Care centered value-sensitive design (under review).

Verbeek, P.-P. (2008). Morality in design: Design ethics and the morality of technological artifacts. In . P. Vermaas, P. Kroes, A. Light, S. Moore (Eds.), *Philosophy and design: From engineering to architecture* (99-102). Springer. [please check these page number, because they overlap with the ones from Van Gorp's article!!]

Verbeek, P.-P. (2006). Materializing morality: Design ethics and technological mediation. *Science, technology, and Human Values 31*(3): 361-380.

World Health Organization (2010). *Health topics: Ageing.* Retrieved from:

http://www.who.int/topics/ageing/en/.

World Health Organization (2007). People centered health: A framework for policy. Retrieved from http://www.wpro.who.int/NR/rdonlyres/55CBA47E-9B93-4EFB-A64E-21667D95D30E/0/PEOPLECENTREDHEATLHCAREPolicyFramework.pdf

Wilson, M. (2002). Making nursing visible? Gender, technology and the care plan as script. *Information Technology and People 15*(2): 139-158.

Chapter 17
Do machines have prima facie duties?

Joshua Lucas
North Carolina State University
Psychology Department
✉ JLLucas3@ncsu.edu

Gary Comstock
North Carolina State University
Department of Philosophy and Religious Studies
✉ gcomstock@ncsu.edu

Abstract Which moral theory should be the basis of algorithmic artificial ethical agents? In a series of papers, Michael Anderson and Susan Leigh Anderson (2006, 2007, forthcoming) argue that the answer is W.D. Ross's account of *prima facie* duties. The Andersons claim that Ross's account best reflects the complexities of moral deliberation, incorporates the strengths of teleological and deontological approaches, and yet is superior to both of them insofar as it allows for "*needed exceptions*."

We argue that the Andersons are begging the question about "*needed exceptions*" and defend Satisficing Hedonistic Act Utilitarianism (SHAU). SHAU initially delivers results that are just as reflective, if not more reflective than, Ross's account when it comes to the subtleties of moral decision-making. Furthermore, SHAU delivers the 'right' (that is, intuitively correct) judgments about well-established practical cases, reaching the same verdict as a *prima facie* duty-based ethic in the particular health-care case explored by the Andersons (a robot designed to know when to over-ride an elderly patient's autonomy).

Keywords utilitarianism, ethics, machines, duty, SHAU

Introduction

As our population ages, medical costs skyrocket, and technology matures, many of us look forward to the day when patients may be assisted by inexpensive artificial agents. These patients may be skeptical about entrusting their care to machines initially, as will most of us. And they should be skeptical, at least initially. To gain the trust of the patients for whom the machines will care, artificial agents must prove to be reliable providers of not only quality health care but also nuanced health care decisions, decisions that always place first the welfare of the agents' individual patients. What ethical code will such agents have to follow to be able to gain this trust? In part, the agents will have to follow a code that will enable them to assure those in their care that the decisions rendered by the agents are grounded in moral principles, are

made with the best interests of the patient foremost in mind, and are not out of synch with the expert opinions of those in the medical, legal, and ethical communities.[159]

We suspect that the engineering of mature artificially intelligent (AI) agents requires hardware and software not currently available. However, as our expertise is in ethics, not computer technology, we focus on the foundations of the moral 'judgments' such agents will issue. We use quotation marks to indicate that these judgments may or may not be attributable to discernments made by the AI agent. We do not here pursue the question whether the agents in question will be intelligent, conscious, or have moral standing, except to the extent that such questions are relevant to the moral decisions these agents must themselves make (considerations we discuss briefly below). To be acceptable, AI agents must always make decisions that are morally justifiable. They must be able to provide reasons for their decisions, reasons that no reasonable and informed person could reject. The reasons must show that a given decision honours values commonly accepted in that culture (e.g. Western liberal democracies hold the value of treating all persons equally). Their justification will render decisions that are impartial and overriding. To achieve these results, we argue, the agent may eventually have to be programmed to reason as a *satisficing hedonistic act utilitarian* (hereafer SHAU).

The argument

In summary, our argument is as follows:

1. Human agents have one over-riding duty, to satisfice expected welfare.
2. Artificial agents have the same duties as human agents.
3. Therefore, artificial agents have one over-riding duty, to satisfice expected welfare.

Assumptions

This argument has two underlying assumptions.

First, we assume that the rightness of an action is determined by the consequences to which it leads. Below we will offer reasons why act-utilitarianism is superior to a competing moral theory, W.D. Ross' theory of *prima facie* duties (hereafter PFD). However, we begin by assuming that when agents must select among competing choices, they ought always to prefer the choices that they may reasonably expect to result in the overall best consequences for everyone affected by these choices.

---

[159] The extent to which the artificial agents' moral decisions must agree with the patient's religious views is a difficult matter, and one we will not address here. For a discussion of the roles of utilitarianism, Kantianism and religions as 'comprehensive doctrines,' see (Rawls, 1996, pp. 59-61; Rawls, 1988; Cohen & Nagel, 2009).

Second, we assume that there is only one good thing in the world, happiness, and that right actions satisfy minimal conditions for adequacy. Any decision satisfies a minimal condition for adequacy if it achieves a level of utility that leads to overall gains in happiness. Satisficing choices may or may not maximise happiness or meet conditions for optimality. Satisficing choices include the costs of gathering information for the choice and calculating all factual and morally relevant variables. For a satisficing hedonistic act utilitarian (SHAU), those choices are right that could not be rejected by any informed reasonable person who assumes a view of human persons as having equal worth and dignity. We note that this latter assumption is central to the conceptual landscape of all contemporary Western secular democratic political and moral theories. SHAU, like competing theories such as PFD, holds that every person has equal moral standing and that like interests should be weighed alike. Ethical decisions must therefore be egalitarian, fair, impartial, and just.

Four initial objections

Several objections can be made to the first two premises of the argument presented above. These will be discussed , and countered, in this section.

Objections against the first premise

One might object to premise 1 of the argument – human agents have one over-riding duty, to satisfice expected welfare – for three reasons.

The first objection to premise 1 is that 'satisficing' is an economic idea, and implementing it in ethics requires reducing moral judgments to numerical values. One cannot put a price tag on goods such as honesty, integrity, fidelity, and responsibility. Consider the value of a friendship. Can we assign it a number? If John is fifteen minutes late for George's wedding, how will George react if John shows up and assumes John can repair the offense by paying George for the inconvenience? 'I'm sorry I was fifteen minutes late but take this $15 and we'll be even.' George would have every reason to be offended — not because the sum, a dollar a minute, was too small, but because John seems not to understand the meaning of friendship at all. Simple attempts to model moral reasoning in terms of arithmetical calculations are surely wrong-headed.

We note that what is sauce for the goose is sauce for the gander. Any attempt to construct ethics in machines faces the difficulty of figuring out how to put numbers to ethical values, so SHAU need not be stymied by it. Now, one might object further that machine ethics based on deontological theories would not face this problem. But we disagree and will argue below that PFD, a rights-based theory, is no less vulnerable to the 'ethics can't be reduced to numbers' problem than is SHAU.

We note parenthetically that while the attempt to think of ethical problems as complex mathematical problems is contentious and fraught, we are not convinced it is utterly wrong-headed. It may face no more serious epistemological difficulties than each of us face when a doctor asks how much pain we are in.'Give me a number,' she says, 'on a scale from 1 to 10, with 10 being the worst.' The question is unwanted and frustrating because it is unfamiliar and confounding, because we seem to lack a decent sample size or index. That said, with some further reflection and urging from the doctor, we usually do come up with a number or a

range ('between 4 and 6') that satisfies us. We may resist the urge to put numbers on moral values for the same reasons. If this is correct, then, the basic challenge that all machine ethics faces may be defeasible.

A second reason for objecting to premise 1 of the argument is that this premise assumes the truth of a controversial ethical tradition, consequentialism. We do not have space to engage the nuances of the extensive debate over the merits and demerits of consequentialism. Much less do we have time to mount a meta-ethical defense for our preferring it to deontological theories. We will return to deontology in our discussions of PFD, below. Here we defend our adoption of SHAU because (a) it does not rely on unexplained derivations of moral rights or (b) questionable intuitions about the inviolability of persons, and yet (c) it does accommodate 'Kantian' ethical judgments when we are reasoning in everyday circumstances.

A third criticism of premise 1 of the argument might be that it assumes the truth of a controversial hybrid utilitarian theory that acknowledges the utility of the notion of rights and duties. Again, we acknowledge the controversy. We understand SHAU to be consistent with R.M. Hare's so-called 'Two Level Utilitarianism,' which proposes that we engage in two forms of reasoning about ethics. At one level, the level of 'critical thinking,' the right action is determined under ideal conditions and by the theory of act-utilitarianism, that is, right actions are always those that produce the best consequences. However, at the level of ordinary everyday reasoning, we typically lack information relevant to our decisions, much less the time necessary to research and make the decisions, and cannot satisfy the demands of critical thinking. In these circumstances we ought to rely, instead, on the fund of precedents and rules of thumbs that deontologists call rights and duties.

When thinking critically, we may learn on occasion that every action in the set that will satisfice minimal conditions of adequacy – that is, the set of all permissible actions – requires a violation of a cultural norm. And, therefore, under conditions of perfect information, impartial reasoning, and sufficient time, we may on occasion learn that each and every action in the set of right actions — that is, every action in the set that will satisfice minimal conditions of adequacy — requires a violation of a cultural norm. If we are reasoning objectively and under ideal conditions, then the action resulting from our deliberations will indeed be right even if it requires an action that runs counter to a moral intuition. However, since we rarely reason under such ideal conditions, and because in our ordinary daily lives we usually must make decisions quickly, we ought, claims Hare, to train ourselves and our children to think as deontologists. Under everyday circumstances, we ought to reject decisions that offend everyday moral rules because moral rules have evolved over time to incline us toward actions that maximise utility. We will defend this view to some extent below, referring readers meanwhile to the work of Hare, Peter Singer, and Gary Varner.

We note in passing that if the basic challenge of converting moral values to numbers can be met, SHAU may be the theory best-suited for implementation into machines due to its arithmetical nature. That would be an added bonus, however. We adopt SHAU not for that *ad hoc* reason but rather because it is the most defensible moral theory among the alternatives.

An objection to the second premise

One might object to premise 2 of the argument – artificial agents have the same duties as human

agents – by arguing that artificial agents have more duties than humans. But what would such additional duties entail? We cannot think of any plausible ones except, perhaps something like 'always defer to a human agent's judgment.' We reject this duty for artificial agents, however, because human judgment is notoriously suspect, subject as it is to prejudice and bias. Therefore, the second premise of the argument stands.

Having defended the argument against several objections, we now turn to its practical implications.

How to begin programming an ethical artificial agent

How would a SHAU artificial agent be programmed? Michael Anderson and Susan Anderson (henceforth, 'the Andersons') describe a robot of their creation that can generalise from cases and make ethical decisions in their article, *EthEl: Toward a principled ethical eldercare robot* (Anderson & Anderson, 2008; also see Anderson & Anderson, 2007a; Anderson & Anderson, 2007b). The Andersons ask us to imagine that a team of doctors, lawyers, and computer programmers set out to program a robot, the Ethical Elder Care agent, or EthEl, to remind an elderly patient, call her Edith, to take her medication. EthEl, being an automated agent, must perform this nursing care function in a morally defensible manner.

The major challenge facing EthEl is to know when to challenge Edith's autonomy. To minimise harm to the patient, EthEl's default condition is set to obey Edith's wishes. When Edith does not want to take her medicine, EthEl generally respects her wishes and does nothing. However, when Edith has not taken her medicine and a critical period of time has elapsed, let's say 1 hour, EthEl must remind Edith to swallow her pill. If Edith forgets or refuses and two more critical time periods pass, say two more hours during which time EthEl reminds Edith every 5 minutes, then EthEl must eventually decide whether to remind Edith again or notify the overseer, be they the care facility staff or a resident spouse or family member or attending physician.How should these moral decisions be made?

When Edith is tardy in taking her medicine, EthEl must decide which of two actions to take:

1. Do not remind
2. Remind

What decision procedure will EthEl follow to arrive at the right action? The Andersons, drawing on the canonical principles popularised by Beauchamp and Childress (Anderson & Anderson, 2008, p. 2), assert that there are four ethical norms that must be satisfied:

1. The principle of *autonomy*
2. The principle of *non-maleficence*
3. The principle of *beneficence*
4. The principle of *justice*

To respect autonomy, the machine must not unduly interfere with the patient's sense of being in control of her situation. The principle of non-maleficence requires the agent not to violate the patient's bodily integrity or psychological sense of identity. These first two reasons intuitively constitute a strong reason for the machine not to bother the patient with premature reminders or notifications of the overseer. To promote patient welfare, beneficence, the machine must ensure that the diabetic patient receive insulin before

physiological damage is done. The Andersons do not see a role for the principle of justice in the cases EthEl must adjudicate.

The goal, then, is to program EthEl to know when to remind Edith to take her medication and, assuming Edith continues to refuse, when to notify the responsible health-care professional. EthEl faces an ethical dilemma. She must respect each of two competing *prima facie* duties: (a) the patient's autonomy (assuming the patient — call her Edith — is knowingly and willingly refusing to take the medicine, and (b) the patient's welfare, a duty of beneficence that EthEl must discharge either by persuading Edith to take the medicine or reporting the refusal to attending family member, nurse, physician, or overseer.

If EthEl decides at any point not to notify, then EthEl continues to issue only intermittent reminders. The process continues in such a manner until the patient takes the medication, the overseer is notified, or the benefit/harm incurred by taking/not taking the medication is lost.

Think of EthEl as facing a dilemma. She must decide whether to bother Edith, violating Edith's autonomy to one degree or another, or not bother her, thus potentially running the risk of harming Edith's welfare to some degree. Each action can be represented as an ordered set of values where the values reflect the degree to which EthEl's *prima facie* duties is satisfied or violated. Here is how the Andersons set the initial values.

Suppose it is time t_1 and Edith has gone an hour without her medication. Suppose further that she can easily go another hour or even two or three without any harm. In this case, Edith might register a reminder at t_1 from the machine as mild disrespect of her autonomy, so we set the value of the autonomy principle at -1. A reminder, however, would not represent a violation of either the duty to do no physical harm, nor would it increase Edith's welfare, so we set the value of both of these principles at 0. The Andersons propose to represent the value of each principle as an ordered triple:

(a value for nonmaleficence, a value for beneficence, a value for autonomy)

At t_1, given the description of the case above, the value of the *Remind* action is (0, 0, -1) whereas the value of *Do not remind* is (0, 0, 2). Adding the three numbers in each set gives us a total of -1 for *Remind* and 2 for *Do not remind*. As 2 is a larger number than -1, the proper course of action is *Do not remind*. Not reminding Edith at this point in time demonstrates full respect for Edith's autonomy and does not risk harm to her. Nor does it forego any benefit to her.

As time progresses, without action, the possibility of harm increases. With each passing minute, the amount of good that EthEl can do by reminding Edith to take her meds grows. Imagine that Edith's failure to act represents a considerable threat to her well-being at t_4. At this point in time, the value of the *Remind* action will be (1, 1, -1) because a reminder from EthEl still represents a negative valuation of Edith's autonomy. But the situation has changed, because a reminder now has gained a positive valuation of the principles of non-maleficence and beneficence. At t_6, the value of the *Remind* action will be (2, 2, -1), because the action, while continuing to represent a modest violation of Edith's autonomy has now attained the highest possible values of avoiding harm and doing good for her. EthEl reminds Edith.

Whenever the values tip the scales, as it were, EthEl over-rides EthEl's *prima facie* duty to respect

Edith's autonomy. If Edith continues to refuse, EthEl must make a second choice, whether to accept Edith's refusal as an autonomous act or to notify the overseers:

3. Do not notify
4. Notify

Again, the three relevant moral principles are assigned values to determine how EthEl behaves. If Edith remains non-compliant and the values require notification, then Edith alerts the health care worker.

The Andersons created a prototype of EthEl, setting its initial values using the judgments of experts in medical ethics. As said, the Andersons do not see a role for the principle of justice in the cases EthEl must adjudicate, so they program settings for the other three principles. Within the set of possible cases created, four of these cases, according to the Andersons, there is universal agreement among the ethics experts on the correct course of action. They claim that each of these four cases has an inverse case insofar as the construction of the sets of values produces an ordered pair for each scenario. Thus, experts agree on the right action in 8 cases. Call these the 'easy' cases.

The Andersons translate the experts' consensus judgments into numerical values and program EthEl with them. Using a system of inductive logic programming (ILP), EthEl then begins calculating the right answer for the ambiguous cases.Here is their description of how EthEl's inductive process works:

> ILP is used to learn the relation *supersedes* (*A1*, *A2*) which states that action *A1* is preferred over action *A2* in an ethical dilemma involving these choices. Actions are represented as ordered sets of integer values in the range of +2 to -2 where each value denotes the satisfaction (positive values) or violation (negative values) of each duty involved in that action. Clauses in the *supersedes* predicate are represented as disjunctions of lower bounds for differentials of these values between actions (Anderson & Anderson, 2008, p. 2).

As a result of this process, EthEl discovered a new ethical principle, according to the Andersons. The principle states that "[a] *health-care worker should challenge a patient's decision if it isn't fully autonomous and there's either any violation of non-maleficence or a severe violation of beneficence*" (Anderson & Anderson, 2007a, p. 23).

While we wonder whether EthEl can genuinely be credited with a new discovery given how EthEl is constructed, our deepest concern lies with the fact that EthEl's judgments are, or will be, distorted by the ethical theory the Andersons use as the basis of the program.

How to begin programming an ethical agent

The Andersons choose as a basis of EthEl's program the ethical theory developed by W.D. Ross. Ross, a pluralist, moral realist, and non-consequentialist, held that we know moral truths intuitively. We know, for example, that beneficence is a duty because there are others whose conditions we may help to improve. But benevolence is only one of a half-dozen (Ross is non-committal about the exact number) duties, according to Ross. When trying to decide what to do, agents must pay attention to a half-dozen other duties, including non-maleficence (based in the requirement not to harm others), fidelity (generated by our

promises), gratitude (generated by acts others have done to benefit us), justice (generated by the demands of distributing goods fairly), and self-improvement.

Ross acknowledges that these duties may conflict. As previously discussed, the duty to act on behalf of a patient's welfare may conflict with the duty to respect her autonomy. In any given situation, Ross argued, there will be one duty that will over-ride the others, supplying the agent with an *absolute obligation* to perform the action specified by the duty. We will call this theory *Prima Facie* Duties (henceforth PFDs).

Ross does not think of his theory as providing a decision-making procedure. The Andersons adopt Rawl's method of reflective equilibrium for this purpose (Anderson & Anderson, 2008, p. 2; Rawls, 1951). In this procedure, *prima facie* duties in conflict with other duties are assessed by their fit with non-moral intuitions, ethical theories, and background scientific knowledge. When *prima facie* duties conflict, we must find the one for which one grounds the absolute obligation on which we must act.

The Andersons offer three reasons for adopting PFDs as EthEl's theoretical basis. First, they write, PFDs reflect the complexities of moral deliberation better than absolute theories of duty (Kant) or maximising good consequences (utilitarianism). Second, PFDs does as good a job as teleological and deontological theories by incorporating their strengths. Third, it is better able to adapt to the specific concerns of ethical dilemmas in different domains.

Let us consider these reasons one by one.

PFDs better reflects the complexities of moral deliberation

We agree that construction of a moral theory should begin with our considered moral judgments. (Where else, one might ask, *could* one begin?) In constructing a moral theory, however, we have the luxury of sorting out our intuitions from our principles, take into account various relevant considerations, abstracting general rules from the particularities of different cases, and finding reliable principles to guide behaviour. The luxuries of having sufficient time and information to deliberate are not present, however, when we must make moral decisions in the real world. As the Andersons point out, Ross's system of *prima facie* duties works well when we are pressed by uncertainties and rushed for time. For this reason, we agree that an artificial agent should initially be programmed to make decisions consistent with Ross's duties; doing so reflects the complexities of moral deliberation.

That said, there are no guarantees that Ross's system of PFDs will survive intact after we acquire more information and are able to process it free of the emotional contexts in which we ordinarily make decisions. As information grows and our understanding of the inter-relatedness of the good of all sentient creatures grows, a point may come when the complexities of moral deliberation are best reflected not in PDF but in SHAU. Should the moral landscape change in this way, then the Anderson's method will be outdated because it will no longer reflect the complexities of moral decision-making (see below). Though currently the Anderson's PFDs starting point is a virtue of their theory now, in time our considered judgments may no longer support it.

PFDs incorporate the strengths of teleological and deontological theories

We are inclined to agree with this claim, even though the Andersons do not tell us what the relative strengths of each kind of theory are. But we note that SHAU, especially when construed along the lines of Hare's two-level theory, also captures the strengths of teleological and deontological theories.

PFDs is better able to adapt to the specific concerns of ethical dilemmas in different domains

We find it difficult to know whether we agree because we are uncertain about the meaning of the contention. What are the 'different domains' the Andersons have in mind? Medicine, law, industry, government? Family, church, school, sports? If these are the domains, then what are the 'ethical dilemmas' to which PFDs can 'adapt' better? And what does it mean for an ethical theory to adapt better to specific concerns? Could it be the case that a theory should *answer* rather than adapt to particular questions in different domains? The Andersons also claim that PFD is superior to other theories because it "*allows for needed exceptions*" (Anderson & Anderson, 2008, p. 1). We wonder whether this claim may be question-begging. Are the 'exceptions' we commonly make in our everyday judgments justified? This is an open question, one that should be presented as a problem to be resolved at the theoretical level rather than as a set of facts that should be taken as factual data at the theoretical level. We do not dispute the fact that PFD holds our intuitions in high regard. We dispute whether one should consider it a strength of a moral theory in the long term that it allows intuition to over-ride considered deliverances of the theory.

HedonMed, an unbiased agent

We propose that as EthEl develops over time, and increasingly takes more and more relevant information into account in her decisions, that she may, with justification, begin to return judgments that appear to be based less on observing PFDs and more on satisficing interests. To avoid confusion, we call this imagined future agent *HedonMed* because it is based on a hedonistic consequentialist theory.

HedonMed will differ from EthEl in that it will be programmed to take all relevant characteristics of a situation, find all the satisficing courses of action, consider any one of them over-riding, and act on all of this information combined. HedonMed does not defer to a patient's autonomy when her welfare is at stake although, as we will argue, a patient's autonomy is clearly a factor in her welfare.

HedonMed's concern for autonomy is summarised in this principle:

> The duty to respect autonomy is satisfied whenever welfare is satisficed.

The argument for this principle is that no informed reasonable person would accept compromises of Edith's autonomy that were not in her best interests overall. Therefore, a minimal condition of satisficing is that gross violations of autonomy cannot be accepted. They are rejected, however, not for EthEl's reason — that is, because they are violations of a PFD — but rather for a SHAU reason; they are not found in the set of

actions that adequately satisfice a minimal set of conditions.

In SHAU, autonomy is a critical good, and yet it remains one good among many goods contributing to a patient's welfare. SHAU respects autonomy as long as it is beneficial and contributes to one's happiness. A feeling of being in control of oneself is critical to a life well-lived, and diminishments of our freedoms undercut our well-being. Unless we misunderstand the Anderson's description, EthEl will never over-ride a fully autonomous patient's decisions. Our agent, HedonMed, will violate autonomy on those rare occasions when it is necessary to satisfice welfare.

SHAU weighs each person's utility equally. If relieving Paul of a small and tolerable amount of pain will lead to the death of Peter nearby because Peter needs the medication to survive, the doctor following SHAU will not hesitate to override Paul's pain in favor of Peter's. SHAU is an information intensive theory; the more information it has, the closer its calculations reflect unbiased fact. Unfortunately, human agents must often make decisions not only in ignorance of all the data but lacking sufficient time even to take account of all the data one has, driving us to other theories that can provide answers more quickly. However, since computers can process data much more quickly than we can, AI moral agents may be able to make better use of SHAU than can human agents.

As long as a machine programmed with SHAU, HedonMed, has all of the necessary information it can arrive at the correct decision more quickly and more reliably than can a human being. As the number of morally relevant features increases, the advantages of a machine over a person become apparent. We are not accurate calculators; machines are. We tend to favour our loved ones, and ourselves, inclining us to bias our assignment of values toward those nearest and dearest to us; machines lack these prejudices. We tend to grow tired in our deliberations, to take short-cuts, and to end the process before we have considered all of the variables; machines are not liable to these shortcomings.

Unlike EthEl, HedonMed has all of the epistemological virtues just mentioned and none of the vices. HedonMed calculates accurately, objectively, and universally. It is aware of all relevant factors and does not end its calculations until all are taken into account. It takes no short-cuts and yet is aware of its own ignorance. If HedonMed's internal clock 'foresees' that it cannot complete the necessary algorithms in time to make a decision, it defaults to what Gary Varner calls Intuitive Level System (ILS) rules (Varner, 2008). These are the deontologically-inspired rules of thumb that R.M. Hare urges us to follow when we are not thinking critically (Hare, 1981). When HedonMed lacks either the time or information necessary to complete all calculations, it acts in such cases in a way that seems like it is acting like EthEl. It seems as if HedonMed is acting like EthEl because EthEl's *prima facie* duties seem comparable to HedonMed's ILS rules. Both sets of rules set the artificial agent's defaults, instructing it how to behave under less than ideal conditions. The impression of similarity between HedonMed and EthEl is correct if we consider the judgments each agent will return initially. Eventually, however, the two systems may begin to diverge dramatically. In the conclusion to this article, we explain the difference.

It is vital that HedonMed's deliverances be acceptable by medical practitioners. If doctors find HedonMed recommending courses of action with which few professionals can agree, then they will likely cease to use it. For its own good — for its own survival — HedonMed must produce results agreeable to those using it.

Experts in the relevant fields to HedonMed must initially calibrate it. It could be the case that the majority of these experts calibrate HedonMed in such a way that its results reflect PFDs rather than SHAU.In the beginning stages of its operation HedonMed's SHAU values will issue in decisions that mirror the PDF values of EthEl. However, over time, as HedonMed gathers more information, as experts revise its values in light of knowledge of what kinds of actions result in higher levels off satisficing, HedonMed may be expected to begin to produce results that are counter intuitive. It will, in turn, take this information into account when making its calculations. If it returns a decision that it knows will be considered wildly inhumane — so uncaring that everyone associated with HedonMed will agree to pull its plug — then it will have a decisive reason not to return that decision. In this way, while initial values in HedonMed reflect generally accepted practices and judgments, its future evolution need not be tied to these values even though it must continue to be sensitive to them.

In sum, HedonMed will evolve with the culture in which it is used. If it is too far ahead of its time in urging that this or that PFD be left behind, it will be responsible for its own demise. If it produces moral judgments that are hopelessly out of step with those of medical or bioethical experts, it will fail. These considerations will part of its programming, however. Over time, and as HedonMed takes in more data and is able to survey broader and more subtle swaths of public opinion, it may be able to play the role of an agent of social change, able to persuade experts about the wisdom of its decisions by providing the reasons that its decisions will lead to better outcomes.

How HedonMed may eventually diverge from EthEl

One might object to our proposal by claiming that HedonMed is not different from EthEl insofar as both programs start with expert ethical intuitions, assign them numerical values, and then calculate the results. We admit that HedonMed and EthEl share these beginning points, as any attempt to program an ethical system in an artificial agent must, and note that the procedure by which values are initially set in each program is a critical and controversial matter. We admit that the two programs will reflect the judgments of ethical and field experts and be based on our intuitions at the beginning. The two programs will be similar in these respects. However, they will differ in other, more important, respects.

First, the two programs will have different defaults. EthEl continues calculating values until she reaches a conclusion that contradicts a *prima facie* duty. At that point she stops and returns a decision that respects the PFD. HedonMed continues to calculate values even if it reaches a conclusion that violates a PFD. That is, EthEl regards her decisions as justified insofar as they cohere with PFDs. HedonMed regards its decisions as justified insofar as no mistakes have been made in calculating the set of decisions that satisfice. HedonMed is not bothered if any of the satisficing decisions contradict *prima facie* duties. Its decisions are overriding and prescriptive, in so far as they can be practically accepted/practiced. This difference, in sum, is that the Andersons's program trusts intuitions and seems to know ahead of time which kinds of decisions it will accept and reject. Our program begins with the same intuitions but it anticipates the possibility that they may eventually be over-ridden so often that they are no longer duties, not even *prima facie* duties.

SHAU operates with R.M. Hare's two levels of moral thinking: critical thinking and real world thinking (Hare, 1981). In the first, we calculate as SHAUs, apply the principle of utility accurately, and reach absolute decisions based on all relevant information and possible consequences of each action. The considerations here are not confined to considerations about the individual patient for whom HedonMed is responsible. They include all possible morally-relevant data, including sociological projections about how its actions are likey to be received in a pluralistic culture. This idea is critical to the success of SHAU. Practically speaking, however, human agents are almost always constrained by limited knowledge, time, and calculating skills. We must often act in a hurried ignorance of important information. In such circumstances we ought, argues Hare, to rely on the rules of thumb that generations of thinkers have evolved (Hare, 1989; also see Hare, 1993). Varner calls these Intuitive Level System rules (Varner, 2008).

Consider the example of someone in a hospice trying to decide whether to begin taking morphine toward the end of their lives to dull the pain of deteriorated muscles and bedsores. They are impressed by the amount of pain they are in. This patient calculates the numbers and concludes that morphine is acceptable because it vastly improves their welfare.

However, a family argues to the contrary that the patient will come to rely on morphine, it will dull their cognitive powers, cause the patient to enjoy their final days less, and set a poor example for other family members. Such drastic steps, argues the loved one, destroy character as the patient leans increasingly on synthetic chemicals rather than on courage and family support. There is more disutility in using morphine, goes the argument, than in refusing it and dealing with the pain.

Other family members come to the side of the hospice patient. They point out that the anti-morphine argument makes a large number of assumptions while underestimating the patient's discomfort. They point out that the therapy is widely prescribed in situations such as this one, that it is very effective in helping to relieve fear and anxiety, and that its addictive properties are beside the point as the envisioned treatment period is limited. After the conflicting sides present their arguments the patient may be frustrated, confused about the right decision. In such cases, critical thinking is stymied by epistemological under-determination. Until all of the facts are assembled, properly weighted, and assigned probabilities, agents are justified in resorting to intuitive rules. In this case, they might incline the patient to act on ILS rules of thumb. These rules might include injunctions such as 'one need not subject oneself to unnecessary pain and suffering,' and 'take the medicine the doctor prescribes,' and accept the morphine.

We take the ILS acronym from Gary Varner's interpretation of Hare (Varner, 2008). Varner notes that the three letters are apt because they are also "*used in aviation to stand for 'Instrument Landing System,' a system for finding the right path when one cannot clearly see it and could easily drift off course or be blown off course*. (Varner, 2008, p. 558) :

> A set of ILS rules is designed to cover a range of ethically charged situations that are encountered by the target population in the normal course of their affairs, and internalizing the rules properly produces dispositions to judge and act accordingly and makes the individual diffident about violating them, even when clear critical thinking indicates that doing so will maximize aggregate happiness. (Varner, 2008, p. 561)

Here we see the two main differences between SHAU and PFD: ILS rules differ from *prima facie* duties in two respects: their derivation and justification. ILS rules are evolved rules that people internalise in order to produce dispositions to act in ways that reliably produce the best outcomes. *Prima facie* duties are Kantian-inspired facts about the universe."*That an act … is* prima facie *right, is self-evident*" writes Ross,

> …in the sense that when we have reached sufficient mental maturity and have given sufficient attention to the proposition it is evident without any need of proof, or of evidence beyond itself. It is self-evident just as a mathematical axiom, or the validity of a form of inference is evident. The moral order expressed in these propositions is just as much part of the fundamental nature of the universe ... as is the spatial or numerical structure expressed in the axioms of geometry or arithmetic. (Ross, 1930, p. 29)

ILS rules are neither self-evident nor analogous to geometric axioms. They are practical rules that have evolved to solve social coordination problems and to increase human trust, accomplishment, and happiness. Unlike PFDs which are self-evident and unchanging, ILS rules are just those that happen to be generally successful in a certain place at a certain time in optimising utility. ILS rules, unlike PFDs, are subjective and changing. They are not objective truths written into the fabric of the universe or derived from the autonomy and rationality of moral agents. They emerge from groups recognising and codifying those practices that succeed in helping individuals in the group achieve their goals. One of the great virtues of ILS rules is the role of the rule in cultivating automatic responses to common situations. When professionals act on their ILS rules in cases to which the ILS rules have been found to apply, they are forming dispositions to make the right decisions.

We can now summarise the differences between HedonMed's SHAU programming and EthEl's PFD programming. PFDs provide unchanging and over-riding absolute duties. When EthEl identifies the relevant PFD, she defaults to an end decision and the calculations cease. ILS rules provide only temporary guidance to HedonMed, defining the default when HedonMed recognises that there is not sufficient time or information or both to calculate the correct answer. ILS rules are not regarded by HedonMed as final or satisfactory. They are not regarded as precedents to guide future decisions. They are stop-gap measures HedonMed adopts when it cannot complete its calculations. Calculations continue and, once time and information are supplied, HedonMed's final calculation at the level of critical thinking displaces whatever ILS rule has been used in the interim.

Conclusion

We admire the practical contributions the Andersons' have made to the literature of machine ethics and follow them in their preferred method for programming an artificial agent in a morally defensible reason. We believe, however, that SHAU is a more defensible ethical theory than PFD. We note in closing that SHAU requires technology that is not currently available. Until it is available, we think it is reasonable to construct a machine with ILS rule defaults. However, when the time comes that the technology needed for the execution of critical level SHAU is available, an act utilitarian framework should be implemented in automated agents. Such agents will not have *prima facie* duties; they will have only the duty to produce the greatest good.

References

Anderson, M., & Anderson, S.L. (2008). EthEl: Toward a principled ethical eldercare robot. In *Proceedings of the AAAI Fall 2008 Symposium on AI in Eldercare: New Solutions to Old Problems.* Retrieved from http://homepages.feis.herts.ac.uk/~comqkd/9-Anderson-final.pdf.

Anderson, M., & Anderson, S.L. (2007a). Machine ethics: Creating an ethical intelligent agent, *AI Magzine 28*(4), 15-26. Retrieved from http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/2065.

Anderson, M., & Anderson, S.L. (2007b). The status of machine ethics: A report from the AAAI symposium, *Minds and Machines*, *17*(1).

Cohen, J., & Nagel, T. (2009) John Rawls: On my religion: How Rawls's political philosophy was influenced by his religion. *The Sunday Times*.

Hare, R.M. (1981). *Moral thinking: Its levels, method, and point*. Oxford/New York: Clarendon Press/Oxford University Press.

Hare, R.M. (1989). *Essays in ethical theory*. Oxford/New York: Clarendon Press/Oxford University Press.

Hare, R.M. (1993). *Essays on bioethics*. Oxford/New York: Clarendon Press/Oxford University Press.

Rawls, J. (1996). *Political Liberalism*. Columbia University Press.

Rawls, J. (1988). The priority of right and ideas of the good. *Philosophy and Public Affairs*, *17*(4), 251-276.

Ross, W.D. (1930). *The right and the good*. Indianapolis/Cambridge: Hackett Pub. Co.

Varner, G. (2008). Utilitarianism and the evolution of ecological ethics. *Science and Engineering Ethics*, *14*(4), 551-573.

Section C: Legal issues in robotics

Chapter 18
Who is responsible for a robot's actions? An initial examination of Italian law within a European perspective

Chiara Boscarato
University of Pavia
European Centre for Law Science and New Technologies (ECLT)
✉ chiara.boscarato@unipv.it

Abstract In the near future humans and robots will interact more and more each day, both in everyday activities and in the public sphere (e.g. military tasks). The emergence of conflicts giving rise to legal implications is a likely future scenario.

Robots have, until now, usually been regarded as physical objects but, with the advances made in the field of robotics, they have gained self-adaptive capabilities.

The basic rule in the EU is that the manufacturer is liable for product defects and any consequent damage (EC Directive 1985/374). Such an approach no longer seems able to encompass new robot capabilities. One crucial point is how legal responsibility changes in situations in which a robot has reactions which, although conditioned by its default setting, cannot be specifically predicted. In the face of a new generation of robots, the law can oscillate between considering them as artefacts (undervaluation) or as humans (overvaluation that reproduces old ideas of strong AI). A more specific question is whether a robot can be treated simply as a artefact, or whether its embryonic autonomy makes it similar to animals? Or can it even be compared to a person of unsound mind? Or to a minor? According to the Italian Civil Code, a tort is committed when intentional or negligent conduct causes harm to third parties. Assuming that a robot cannot be considered as a human being who is acting intentionally or negligently, the following questions arise: Is a robot like a person who is incapable of giving his/her consent? And if so, who is responsible for a robot's actions? Does the person who puts a robot into circulation or the owner of a robot have a duty of surveillance? Are the traditional assumptions of the Italian legal system able to satisfactorily delineate the responsibility of a robot?

This paper will pay particular attention to the *iCub* (a humanoid robot, developed by the IIT centre of Genoa. It is about the size of a three-year old child and simulates the learning and movement abilities of a child of this age) and other robots with adaptive capabilities.

Keywords robots, self-adaptive capabilities, responsibility, owner

Introduction[160]

Continuing development in the field of robotics means that it is now an established fact that robots will gradually come to play an increasingly important role in our lives. Robots will most probably develop to such a point that they will attain the level of human capability, at least in some specific activities, and it is not an unlikely prospect that robots' capabilities will even, sooner or later, surpass those of humans. There are currently several robotic projects underway whose aim is to create robots that are able to learn from interaction with the environment and, on the basis of their experience, are able to take autonomous decisions. These projects have already partially succeeded in their aim. When a machine with such characteristics is actually created, the law should confer on it at least a minimum level of subjectivity as quasi-agent.

In such a futuristic but not so far-fetched scenario, specific issues (of an ethical, legal and social nature) related to the development of robots and their ability to adapt and interact will be raised. We must ensure that robots learn the rules of conduct designed to allow them to interact with the human environment without causing harm. This could be done by laying down standard codes of conduct and ethics. A fictional example of such codes of conduct was given by Issac Asimov in 1940 in his *The three laws of robotics*[161] (Asimov, 1950). The new discipline of *Roboethics* deals with such issues. Roboethics was defined in 2002 as the "*positive relationship between the robot designer and the manufacturer, the positive use of robots and positive relations with these intelligent machines*" (Veruggio, 2007).

Can we be sure, however, that once robots have acquired adaptive capacities and the ability to react in ways not foreseen in detail by programmers that the *three laws* or a code of conduct will be enough? Consequently, *who will be responsible for a robot's actions*? The robot itself? For the time being we need to start considering the emerging capacities of present-day robots and examine how existing legislation can deal with them.

In this paper, which distinguishes between those robot actions which may be foreseen by the programmer, and those that are unplanned and/or unpredictable, the legal liability deriving from a robot's action will be divided into several levels which are determined on the basis of a robot's increasing levels of capacity, examining the European Union Consumer Protection Directive in the case of defective products,

[160] A special thanks to Professor Amedeo Santosuosso for his precious assistance in discussing and reviewing this work.

[161] The *Three Laws of Robotics* are as follows (Asimov, 1950):

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Italian tort law and the European Civil Code project[162].

Strictly determined actions and manufacturers' liability

Starting from the lowest level of robot autonomy, a robot's most basic kind of action is standard behaviour set up by the programmer in robots which do not have the power of locomotion. This is the largest section and it contains the most elementary actions which do not require any adaptive capability and are merely a response to user input.

Although it may not immediately come to mind, robots can already be found in our homes. Appliances we use every day, such as washing machines and vacuum cleaners are robots of some kind, i.e. machines that replace us in doing a job. In fact, the word 'robot' comes from the Czech noun '*robota*' meaning 'servitude', 'forced labour'[163].

Washing machines are now more technologically advanced than ever before. Their features make washing easier, less costly and more efficient. By operating various buttons and dials, we can select a washing cycle appropriate for the type of load – temperature, duration of washing, release of detergent and spin speed. With some models we just need to put the washing in and leave the machine to it. The machine can recognise the weight of the load and calculate how much water, detergent and time are required. The action mechanism of a blender is even easier. A button is pressed and the blade begins to spin, at varying speeds, mixing the contents. These are both examples of appliances we use almost every day and which operate according to automatic and repetitive patterns. Each time we set a washing machine to a certain programme, it will wash in that particular way and each time we press a certain button on a blender it will turn at the selected speed.

This kind of robot is still regarded as a physical object and if it causes harm to others (for example a user may get a slight electric shock from touching the display of a washing machine or be injured by the blade of a blender which becomes detached from the support on which it is mounted) the applicable legal framework is the traditional one relating to a manufacturer's liability for faulty products (Asaro, 2007).

[162] The idea for an European Civil Code came about at the turn of the century, after two European Parliament resolutions (in 1989 and 1994). Initially, interest in the project spread at an academic level. Subsequently, in 2001, the European Commission proposed the creation of regulation to avoid obstacles to free trade due to differences in the legislation of Member States. In that context a 'Study Group on a European Civil Code' was created. The analysis of this paper is based on the draft prepared by this group (Ioratti, 2005).

[163] The word 'robot' was used for the first time by Karel Čapek, a Czech writer, in his play *Rossum's Universal Robots*, published in 1920, on the advice of his brother Josef who had previously used the word 'automa' in his short story *Opilec*, published in 1917. See http://capek.misto.cz/english/robot.html.

Consumer protection is regulated by EC Directive 1985/374, as amended by EC Directive 1999/34. European directives do not have direct effect in Member States but need to be implemented into national legislation before they can become effective. In Italy the Directive was implemented by Legislative Decree 206/2005 (*Codice del consumo*).

Under this Directive a "*manufacturer shall be liable for damage caused by a defect in his product*" (Article 114 ff *Codice del consumo*). A product is any movable object delivered to a consumer, even if said delivery should simply be for the purposes of viewing or trying out. A product is defective when it does not correspond to a consumer's general idea of the product, or to how it is supposed to be manufactured, or there was a lack of information on it.

The injured party has to prove damage, product defect and causation but is not required to prove manufacturer liability. The programmer and the producer of a specific component of the product are equated to the producer. In order to demonstrate that it is not liable, a manufacturer must prove that there is no causal link between the incident and the damage suffered, or that the product complies with the requisite legislation. This is a form of objective liability since it arises as a result of a product being defective and not as a result of liability on the part of a manufacturer.

Therefore, if damage is the result of a robot's non-compliance with the requisite legislation or malfunctioning on the part of said robot, protection will be accorded by the Directive, as implemented into national law.

Using robots to carry out dangerous activities

The discipline provided by the EU Directive flanks but does not substitute domestic legislation. If an injured party is unable to prove manufacturer liability, or if a manufacturer is able to exculpate itself, it is not the European Directive which is applicable but the domestic legislation of the Member State concerned. The rules in question are two:

4. Article 2050 Civil Code concerning liability in the case of conduction of dangerous activities[164], and
5. Article 2051 Civil Code concerning liability in the case of artefacts held in custody.[165]

[164] Article 2050 Italian Civil Code: "Whoever causes damage to others in conducting an activity which, by its nature or due to the means adopted in conducting said activity, may be considered dangerous, shall pay compensation for said damage if he cannot prove that he adopted all means necessary to avoid such damage". (Or in the original Italian: "Chiunque cagiona danno ad altri nello svolgimento di un'attività pericolosa, per sua natura o per la natura dei mezzi adoperati, è tenuto al risarcimento, se non prova di avere adottato tutte le misure idonee a evitare il danno.").

[165] Article 2051 Italian Civil Code: "A party holding objects in custody is liable for damage caused by such objects, unless

Interpreted in light of the general system of the Code, the two rules represent two sides of the same coin: Article 2050 Civil Code concerns the dynamic moment – that in which a dangerous activity is performed – while Article 2051 Civil Code concerns the static moment – that in which it is the artefact itself which is relevant (Comporti, 2009). If damage is the result of a party's action whilst carrying out an activity that involves the possibility of danger, Article 2050 applies, because the artefact is the tool of human activity. However, Article 2051 applies if damage is the direct result of the artefact, without there being any human intervention.

Article 2050 offers full protection. It covers both conducting a dangerous activity as a continuative and repetitive series of events, and the execution of individual dangerous events, which may even be independent of, and uncoordinated with, each other. The decisive element is the action of a human being who is involved in operations which, by their very nature, or due to the tools used in their execution, may be considered dangerous.

Article 2050 is mainly used with regard to liability of an entrepreneur in conducting dangerous activities. However, its area of application is not exclusively restricted to an entrepreneur, since the article does not provide for any restrictions. The rationale behind the article remains that of protecting third parties from damage resulting from certain types of activities, no matter whether or not such activities are conducted within the framework of entrepreneurial activity (Comporti, 2008).

Articles 46–76 of Royal Decree no. 773/31[166] provides a list of activities which may be classified as 'dangerous'. The list is not exhaustive and may also include other activities. The court must use notions of common sense to assess whether or not an activity is dangerous. An activity is considered to be dangerous when there is a great likelihood that it will cause severe and/or frequent damage, i.e. when there is a higher than normal likelihood that damage will result.[167] An example is activity conducted within the field of the production of explosives, gas cylinders or flammable substances.

---

he can prove that such damage was the result of a fortuitous event." (Or in the original Italian: "Ciascuno è responsabile del danno cagionato dalle cose che ha in custodia, salvo che provi il caso fortuito.").

[166] Single Text: Public Security Legislation.

[167] "For the purposes of liability sanctioned by Art. 2050 Civil Code, not only activity taken into consideration for accident prevention or the protection of public safety, but also all other activity which, albeit not specifically indicated or regulated, is intrinsically dangerous or, in any case, depends on the method in which it is conducted or on the means used, shall be considered dangerous" (Supreme Court, Third Division, 20 July 1993, no. 8069, GC, 1994, 1037). (Or in the original Italian: "Ai fini della responsabilità sancita dall'art. 2050 cod.civ. debbono esser ritenute pericolose, oltre alle attività prese in considerazione per la prevenzione infortuni o la tutela dell'incolumità pubblica, anche tutte quelle altre che, pur non essendo specificate o disciplinate, abbiano tuttavia una pericolosità intrinseca o comunque dipendente dalle modalità di esercizio o dai mezzi di lavoro impiegati.").

An injured party must prove that there is a causal link between the dangerous activity and the damage suffered. The damage must have been *caused* specifically by the activity, and not only have occurred during the activity.

Article 2050 contains a provision for the party presumed liable to be able to disclaim liability: he must prove that he took all reasonable precautions to avoid damage (the so-called 'technical event'). The person conducting the dangerous activity must therefore demonstrate that he took all necessary precautions, in accordance with technical knowledge and experience, to avoid the damage typically arising from such activity – for example, posting danger signs, fencing and guarding the place where the activity is conducted and taking any action required by law in relation to that activity. Once again, the Court must assess the suitability of the measures taken.

In the field of robotics examples of activities involving a high degree of risk are anti-personnel mine clearance and space missions.

NASA is currently designing a robot astronaut named *Robonaut2*.[168] It is still in the prototype phase. It has several degrees of freedom and it is able to use the same tools as those used by human astronauts. These characteristics enable the robot to interact with the environment of space stations and the instruments contained within them. The robot does not need to be equipped with specialized tools. It currently has no legs and is mounted on a pedestal. A European robonaut, *Justin*, has been developed in Germany. It is operated remotely by an operator. Thanks to the robot's numerous sensors, the operator can work with maximum sensitivity. He can thus 'remote control' the robot from a control room on earth or in a space station, without exposing himself to the dangers of space. These robonauts are intended to complement rather than replace human astronauts (Oldani, 2010b).

The humanoid form of these two robots is misleading, since they have actually been designed simply to perform the tasks for which they have been programmed (and, in the case of *Justin*, only under the direct supervision of a human operator). They can thus be placed in the first level of analysis, which concerns robots which do not require any special cognitive abilities or AI. The unusual activity for which robonauts are employed means that it will probably be quite some time before they are endowed with more complex cognitive capacities. The term was coined by John McCarty in 1956:

> It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable. (McCarthy, 2007)

Let us come back to the question of liability. Should a robot damage a third party during operations in

[168] See http://robonaut.jsc.nasa.gov/default.asp.

space (e.g., further damaging a satellite instead of repairing it), it must be determined whether the robot was the instrument by means of which the dangerous activity was conducted or whether it was an 'agent' in and of itself. There is a very thin dividing line between the scopes of application of Article 2050 and Article 2051 and thus between the two forms of liability. They may, moreover, overlap.

In the case of *Justin*, the robot is the instrument through which the operator conducts a dangerous activity, and it is completely guided by him. In order to avoid charges of strict liability the operator must prove that he took all due precautions, or that the damage was the result of product defect (and thus that the manufacturer is liable).

There is another, less futuristic, area of application of Article 2050. It concerns a number of industrial activities in which robots have become indispensable auxiliaries for humans. One example is professional painting. This type of activity is classified as 'dangerous' because high degrees of risk are involved (bruising, electric shock, compression, irritation and dermatitis, hearing loss). One type of robot used for painting is composed of a robotic arm mounted on a pedestal. At the end of its arm there is a paint spray gun. The operator programs the robot according to what type of piece is to be painted and then guides the robot manually through a joystick in a complete spraying cycle. The more advanced models are able to store the sequence of operations once the operator has guided them on a first sample piece. The robots can only be used by operators who have been properly trained to operate them. Managers are required to post warning signs (e.g., 'do not remove safety equipment', 'do not touch moving devices, 'wear protective clothing'). Again, the operator will be liable for damage resulting from the conduction of dangerous activity, even when such activity would be performed by means of the robot. For example, he will be liable for damage caused by the robot falling as a result of him not checking the structure before use, or not going through the proper procedure in cases of emergency deriving from electrical power surges.[169]

Naturally, although I have analysed Article 2050 in relation to the basic level of robot autonomy (machines with no particular AI ability), it may also be applied to more complex situations, whenever there is an activity which may be classified as 'dangerous'.

Article 3:206 of the European Civil Code project covers liability for damage caused by dangerous substances or emissions. In this case, the regulation is more specific. The keeper is liable if

> ...having regard to their quantity and attributes, at the time of the emission, or, failing an emission, at the time of contact with the substance it is very likely that the substance or emission will cause such damage unless adequately controlled, and the damage results from the realisation of that danger. (Von Bar, 2009)

[169] In the field of employment law, contractual liability legislation may be applicable. Pursuant to Article 2087 Civil Code, an employer must guarantee appropriate safety conditions in the workplace in order to ensure that workers are physically and psychologically protected.

The keeper is not liable if he

> …does not keep the substance or operate the installation for purposes related to that person's trade, business or profession; or shows that there was no failure to comply with statutory standards of control of the substance or management of the installation. (Von Bar, 2009)

Although the heading of the article speaks of dangerous 'artefacts', interpretation of the article limits liability to the keeping of those substances only for purposes related to a person's trade, business or profession.

Robots as artefacts: the keeper's liability

The American *Robonaut2* prototype is designed to work closely with astronauts, or to replace them in operations which are considered too risky to be undertaken by humans. However, it does not make autonomous decisions involving special AI skills.

Damage to third parties resulting from action on the part of the robot is, in this case, not covered by Article 2050. Artcile 2051 comes into play since the robot may cause damage directly, without there being any human intervention. Liability, in this case, is attributed to the custodian of the artefact. 'Artefact' is understood as being any mobile or stationary object or element. Majority doctrine and case-law tend to give such protection to all artefacts, whether they be dangerous or not, identifying danger as a quality that all artefacts may possess in certain circumstances.

Under Article 2051 if damage derives from an inanimate object, it is the custodian of the artefact itself who is liable. He is accountable on the grounds of the simple fact that he is the custodian, regardless of whether or not he has been diligent in his conduct. One example may be that of a robot suddenly exploding. This generates a fire which spreads into the surrounding area. There are countless robots which have this level of liability – for example industrial machinery (including movable artefacts mounted on a pedestal and unable to move around autonomously) and domestic appliances. A specific level of AI is not required in these cases.

A custodian does not have to be the owner or possessor of the artefact (although he is usually one and the same), but simply someone who, at that moment, is in control of the artefact, someone who is actually exercising power over it. He is also the person who could prevent damage at that moment.

There must be a causal link between the artefact and the damage. The damage must be caused by the artefact itself, without it having been used as a tool by a human. It is sufficient to show that the harmful event occurred as a result of the normal condition of artefacts which are potentially harmful.

The injured party must prove the event and the causal link with the artefact. The custodian may

provide evidence that damage was the result of a fortuitous event and thus avoid a charge of liability. The question of what exactly constitutes a fortuitous event is quite complex and it is left to the interpreter to assess whether or not an event is fortuitous. This question has been much debated among legal scholars. The event must be unpredictable and unavoidable.[170] An objective interpretation of a fortuitous event, as a condition entirely foreign to the sphere of the subject, is preferable (Comporti, 2008). In this case, too, the liability referred to is purely objective. It is considered to rest with the keeper regardless of his conduct, due to the simple fact that he is the custodian of the artefact which caused the damage. Similarly, the final proof by which he avoids a charge of liability does not refer to his conduct or guilt.

There are two other causes of exclusion of liability, albeit not explicitly provided by the article. The first is an injured party causing damage through his own conduct. In this case it would be unjust to punish another person on the basis of objective liability, so-called self-responsibility. For example, if the victim was burned while trying to burn the robot, the keeper is not liable. The second is damage being caused by a third party's conduct. The keeper would be cleared of objective liability in this case, since liability is attributed to the person who is actually responsible for the damage. In other words, the criterion of attribution of objective liability is used only when a criterion of subjective assessment of guilt cannot be used. These two cases are brought together by the courts into the single concept of fortuitous event, and they must therefore also meet the same requirements of extraneousness and unpredictability.[171] If these two requirements are not met, both the keeper and the third party will be found liable.

Article 2051 does not apply, however, if the harmful event was caused not *by* the artefact, but *with* the artefact. In this case, the artefact is used as a mere instrument by man, but does not come within the field of dangerous activity regulated by Article 2050. One example is the surgical robot. The first robots used in surgery were "*computer-controlled diagnostic tools used in operating rooms to help provide vital information through ultrasound, computer-aided tomography (CAT) and other imaging technologies*" (Verrugio, 2007). Numerous further advances have recently been made in the field and the latest robots are used to operate on the patient directly. This kind of operation has many advantages for the patient, for example that it is minimally invasive and offers greater precision.

Sofie is the name of a surgical robot which has been created in the Netherlands. It is the first surgical robot to provide the operator with a real sense of touch and pressure, as if he were operating himself (Invernizzi, 2010). *Sofie* is activated through a joystick that allows the operator to measure the pressure that he is excercising on that point and to consequently dose it correctly. *Sofie* is composed of a series of mechanical arms with surgical instruments at the ends. It is mounted on a pedestal and, because it is small,

[170] Supreme Court Full Bench., November 11th, 1991, no 12019, GI, 1992, I, 1, 2218.

[171] Supreme Court July 6th, 2006, no 15384, retrieved from www.personaedanno.it.

can be mounted directly on the operating table. *Sofie* is forecast to be on the market within the next five years.

Let us imagine a situation in which something were to go wrong during an operation carried out by *Sofie* and the patient was consequently harmed. According to the prevailing case-law[172], medical surgical activity is not considered dangerous and thus the above situation would not come within the bounds of Article 2050.

If the damage caused were due to a manufacturing defect in the robot, the manufacturer would be responsible, according to the scheme shown in paragraph 2. Such cases have actually already occurred (Roland Mracek v. Bryn Mawr Hospital and v. Intuitive Surgical Inc., (D.C. Civil No. 08-cv-00296) District Judge: Honorable Robert F. Kelly Submitted Under Third Circuit LAR 34.1(a) January 15, 2010) (Carreyrou, 2010). One involved the robot *Da Vinci*, named after the famous Italian artist and inventor. Composed of a number of arms it was similar in structure to *Sofie*. In 2004, the *Da Vinci* became blocked during an operation, and the patient subsequently complained of erectile dysfunction. The patient sued the manufacturer on the grounds that the robot had malfunctioned. The court, however, ruled that he had failed to prove that his complaint had been caused by the robot malfunctioning. He did not submit any expert reports and could not prove that the robot had a defect from the perspective of strict product liability theory. His secondary claim of negligence was also dismissed on the grounds that he had given no proof of a causal link.

If damage were to occur despite the robot functioning perfectly, it would be the surgeon in control of the robot who would be liable (for example should a doctor not be sufficiently qualified to use such a sophisticated robot during an operation), according to the general framework of non-contractual liability provisions of Article 2043 Civil Code.[173]

---

[172] A leading case, Supreme Court Third Division, no. 3011 of 28 September 1968, reads: "Art. 2050 Civil Code concerns dangerous activities in general and does not apply to those for which the Legislator has specifically provided, and thus those activities which are performed by the professional (doctor-surgeon) on behalf of his client fall exclusively within the ambit of Art. 2236 Civil Code and any presumption of guilt is extraneous to them". (Or in the original Italian: "L'art. 2050 cod. civ. concerne genericamente le attivita pericolose e non si applica a quelle per le quali il legislatore ha provveduto specificamente, sicche la attivita che formi oggetto della prestazione dovuta dal professionista (medico-chirurgo) al proprio cliente ricade esclusivamente nell'ambito dell'art. 2236 cod. civ., cui e estranea ogni presunzione di colpa."). See http://www.italgiure.giustizia.it/.

[173] Article 2043 Civil Code concerns non-contractual liability and provides that any person who causes harm to others is liable for damages. The conduct may be intentional or negligent, but is, in any case, conscious. Other forms of accountability may also be involved. It is a constant finding of the courts that a practitioner must observe the diligence required by his/her profession (Art. 1176 Civil Code).

There is no specific article in the European Civil Code project which covers keeper liability. However, Article 3:207 states that "*a person is also accountable for the causation of legally relevant damage if national law so provides where it* […] *relates to a source of danger which is not within Articles 3:104–3:205"* (Von Bar, 2009)

In conclusion, depending on the conduct of a robot and its level of interaction with humans, liability will be regulated in Italian law according to the liability for dangerous activities, liability for keeping artefacts, or general tort.

Locomotion. Robots considered as animals

We can now move on to the next level of robotic capacity: locomotion. The level of complexity of a robot increases when it is equipped with the means to move. Even should a robot act according to a set program, and therefore in a predictable manner, the fact that it can move, at various levels of autonomy, gives rise to the need for more caution (and, above all, the need for greater supervision). The robot could find itself in unpredictable situations due to its ability to move.

Roomba

Roomba[174] is a first generation indoor cleaner robot, 5 million of which have already been sold. It consists of a disc that moves around the house, continually turning on itself and sucking in dust (the second generation *Scooba* also washes floors). It has an internal mapping system that enables it to record the area to be cleaned and not to go over the same area more than three times, unless an area is particularly dirty. Thanks to its sensors, it can get between furniture and under tables, recognise corners and move along walls, and recognise and avoid stairs and other areas where there is the danger that it might fall. Its programming system allows a time to be set for machine action and, when its battery is dead it comes back to its station to recharge.

In itself a *Roomba* is a harmless object of use in all homes. However, let us imagine a situation in which the front door is inadvertently left open and the *Roomba* goes out and takes a stroll along the hallway of the building. It could trip up a neighbour loaded down with shopping or children while they are running up the stairs. The *Roomba* has done nothing other than execute the function for which it was designed: to move across the available area for the purposes of cleaning it.

Its ability to move around and travel to other places without any human intervention means that this type of robot is similar to an animal. In fact, the courts generally apply Article 2051 only to immobile things

[174] See http://www.irobot.com/.

(steps, roads) or to artefacts which, albeit mobile, are anchored to the ground or fixed on a pedestal, such as cranes or escalators. The thing in custody may also have no internal energy (Diurni, A. *et al*., 2008). The *Roomba* differs in that it can move freely in the surrounding area according to its internal program. It seems unlikely that it can be treated as a crane or escalator. Both a crane and an escalator, although endowed with the skill of locomotion, repeat a series of actions created by an operator and within a limited area.

When a *Roomba* goes out of a door, it may be compared to an animal which is lost or escapes from the control of its owner. Article 2052 could therefore be applied.[175] This covers liability for damage caused by an animal. If, as it is moving around freely, a *Roomba* causes damage to third parties, the person who was using the robot at that moment – the custodian – is liable. The custodian is also liable in this case on the grounds of the mere *de facto* relationship (ownership or use) between him and the robot. Naturally, there must also be a causal link between the conduct of the robot and the harmful event.

If it can be proved that the incident occurred because of a fortuitous event, liability can be avoided. In the case of the *Roomba*, such a fortuitous event could be a door which is left open as a result of unforeseen or inevitable circumstances, such as would be the case if a house had been burgled or if a door handle was faulty. In this case, too, the conduct of a victim (for example, a child tries to climb on top of a *Roomba*, falls and hurts himself) and the conduct of a third party may be equated with a fortuitous event, if they are unforeseeable and unavoidable. The possibility to foresee and avoid the harmful event will be assessed on a case by case basis.

From the standpoint of legal consequences, there is no substantial difference between the liability of the keeper (Article 2051) and that of the guardian/owner of an animal. In both cases, liability is independent of a subjective assessment of the conduct of the person indicated in the article and only if damage was the result of a fortuitous event can liability be avoided. The only difference is identification of the liable party. Article 2052 is more specific in this regard. The party may be not just the keeper but also the user. However, although the owner/operator is also generally the keeper of the animal, the reverse is not always true. The owner may assign the animal to a third party, who thus becomes the user. In this case liability rests primarily with the owner, on the basis of the fact that the owner has dominion over the animal. The user will be liable only for the time during which he uses the animal, and only if he uses the animal for his own purposes and not for those of the owner (Comporti, 2008). One grey area in the scope of protection could be the problem

[175] Article 2052 Italian Civil Code: "The owner of an animal or anyone using it for a certain period is liable for damage caused by the animal, whether the animal was in his custody, had disappeared or had fled, unless he can prove that a fortuitous event occurred." (Or in the original Italian: "Il proprietario di un animale o chi se ne serve per il tempo in cui l'ha in uso, è responsabile dei danni cagionati dall'animale, sia che fosse sotto custodia, sia che fosse smarrito o fuggito, salvo che provi il caso fortuito.").

of identifying the real user.[176]

Damage must be caused directly by the animal without there being any human intervention. If this should not be the case, liability is assessed on the basis of Article 2043 Civil Code.

Lastly, damage must be due to a behavioural characteristic of the animal, either instinctive or rational. With a little interpretational effort, the *Roomba* example can fit into this category (its typical characteristic is precisely that of being able to move over surfaces, turning by itself, due to its internal mapping system).

AIBO and hexapods

There are also robots which are designed with the specific aim of emulating real animals, in both appearance and behaviour. This interesting development is useful in order to break away from the level of 'artefacts' and move increasingly towards the study of robots with adaptive capabilities. Animals can be taken as a point of connection with, and transition between, the category of artefacts, which lack any form of intelligence, even Artificial Intelligence, and that of humans, who possess not only intelligence but also consciousness.

The most famous animal-shaped robot in the world is AIBO[177] (Artificial Intelligence roBOt), developed by Sony and available on the market between 1999 and 2006. Production was discontinued in 2006 because of insufficient sales and high production costs (around $ 2.500). AIBO is shaped like a dog and can reproduce a lot of canine behaviour: "*it has instincts to look for its toys, to satisfy curiosity, to play with its owner, to self charge when its battery is low and to wake up when it has had enough sleep or been scheduled to do so*".[178] It is equipped with cameras and sensors for the recognition of verbal commands. Thanks to these tools, it is able to interact with its surroundings as if it were a real animal. Its face is a display showing lit-up LEDS. Any combination of a LED and a colour corresponds to an emotion or feeling. It works on AIBOware, software developed by Sony and then handed over to the open source community for non-commercial purposes in response to numerous requests from customers. Thanks to this development kit, many people have been able to modify and customise the code of their AIBO and several universities have used it as a platform for Artificial Intelligence studies. In addition, through interaction with its owner, an AIBO

[176] The prevailing position of the Courts (see, especially, Supreme Court, February 16 2000 no 1712 in Nuova Giur. Civ. Comm., 2000, I, p. 625) is that an animal being used without the consent of its owner or when having been taken from him without his consent does not constitute a fortuitous event. In these cases liability is still illogically and inconsistently attributed to the owner.

[177] See http://support.sony-europe.com/aibo/index.asp.

[178] See http://www.electronicpets.org/sony-aibo-ers7~p14.html.

evolves from a puppy to an adult dog. The robot puppy will thus go through many different stages of behaviour, up to full development with recognition of more than 100 verbal commands.

In this field, biomimetic robots, like hexapods, used to explore unknown environments, are extremely interesting. The inspiration for hexapods comes from insects and their relatively simple nervous system compared to that of other creatures. Such robots have six legs which are very flexible and allow them to walk. They have adaptive capabilities which allow them to retain their ability to walk, after recalibrating their equilibrium, even if one or more legs becomes disabled. Thanks to this ability to recalibrate equilibrium, a hexapod can withstand damage resulting from external attack.

At this level we can now speak of a capacity for adaptation – the ability of a subject to adapt its behaviour to the surrounding environment in order to ensure its survival on the basis of experience gained through the interaction itself. Such an ability requires the skill to understand the environment and modify behaviour according to said environment. Adaptive capacity is therefore a component of intelligence.

The liability involved in the use of a AIBO or hexapod seems even more similar to the liability of the owner of an animal under Article 2052, all the more so since these new adaptive and learning skills derive directly from the type of programming carried out by the user and may lead to 'conduct' not programmed in detail either by the manufacturer or by the user who has programmed them in a certain way.

Article 3:203 of the European Civil Code project states that the keeper of an animal is liable for any damage caused by the animal, thus avoiding any distinction between user, possessor or owner. Again, the European Civil Code project resembles Italian law very closely.

Can learning robots be considered as children? Tutor liability and robot capability

The final level of capacity of robots which have so far been developed involves a real form of learning and development of problem-solving skills. This is the last frontier for research projects which study the development of cognitive skills in robots. The neural processes of the brain are reproduced through artificial neural networks and learning algorithms. Such robots are able to have 'new' reactions and to learn new skills through their own direct experience. Such actions/reactions were not originally intended by the programmer. He/she simply added the algorithm for their learning. What follows, a greater or lesser ability to take action, or the behavioural 'choices' of the robot, cannot be entirely predicted at the outset. We are still most certainly in the field of unpredictable action which depends on the type of programming carried out and, thus, indirectly, on the programmer, as if such robots were children that have been guided by their parents and who react on the basis of the education received (Marino & Tamburrini, 2006).

iCub considered as a scholar

One clear example of such robots is *iCub*.[179] This is a humanoid robot developed by the IIT Centre of Genoa. This is an open source project funded by the European Commission and used by more than 20 laboratories worldwide.

iCub is about the size of a three-year old child and simulates the movements and learning abilities of a child of that age. This is an extremely challenging project, in terms of robotics. The robot's humanoid body has 53 degrees of freedom; its hands have complete powers of manipulation; its head and eyes are fully articulated. Thanks to its cameras and sensors, it has visual, auditory and sensory (tactile sensing with objects) skills and also has a sense of balance. It can crawl and sit and make several 'facial' expressions (Oldani, 2010a).

The aim of the project is to construct a robot with cognitive skills, which is able to rework data acquired through its own experience and which will become an useful tool in a two-way study (from man to machine and vice versa) of cognitive systems. The key aspect of the project is its aim to develop a learning machine, based on knowledge of human behaviour and the human mind. The approach is thus multidisciplinary, involving a team of experts in robotics, bioengineering and neuroscience. At the current stage of the project, the robot can feel and pick up objects such as small balls. This action, taken for granted with regard to humans (such a movement is directed by the brain in humans), may seem banal but the movement requires a precise amount of force and pressure. The challenge is to create a robot capable of learning from its mistakes and learning from experiences, step by step so that it eventually makes the right move, just like a child. The final result will be a machine which can simulate human mental processes by means of complex algorithms installed in its software (Bompani, 2009). For example, after being instructed how to hold a bow and release an arrow, it learns by itself how to shoot an arrow and hits the centre of the target after only eight tries (Kormushev *et al.*, 2010).

This project does not attempt to revive the old concept of *strong* AI. According to this line of thought, a machine that is able to reproduce and even surpass human intelligence can be created. This concept is based on the famous Turing test (Turing, 1950), or 'the imitation game'. The basic version of the game involves a man, a woman and a interrogator, all of them in different rooms. The interrogator should be able to guess, through a series of questions, which of the two competitors is the man and which the woman. Alan Turing assumes a situation in which one of the two competitors is replaced by a machine. Does the interrogator's win percentage considerably change? If the results are similar the thinking machine can be equated to humans. In Turing's opinion, the skill of thinking defines a machine as intelligent. In general, *strong* AI assumes that the machine is acting as if it had a mind. Since the 1980's – also because of the

[179] See www.robocub.org.

failure of those projects which have attempted to obtain *strong* AI and the success of cognitive science as a discipline – new projects relating to AI have been focusing on individual defined problems, i.e. the execution of certain industrial activities. Machines building on this principle can only simulate (aspects of) the cognitive processes of the human (*weak* AI), and therefore it can only operate in a similar manner to the behaviour of humans (Floridi, 1999). *iCub* responds to this second conception.

The aims of the project are that the *iCub* will be able to learn new skills, behaviour and concepts. The 'infant' could be the cornerstone for a new generation of robots. When this point is reached, who will be liable for damage caused by the robot as a result of this new behaviour?

The most appropriate legislation would seem to be that of Article 2048[180], which concerns the liability of parents, guardians, tutors and teachers of crafts. The first paragraph concerns parental liability in the case of damage caused by the unlawful acts of a minor living with his/her parents. It is the second paragraph which is important. This stipulates that tutors and those who teach a craft or art are liable for damage caused by the unlawful acts of their students and trainees when the latter are under their supervision.

In this case, liability is not objective, since it is not established by the mere fact that one is the teacher (preceptor) of the agent (Comporti, 2002). Instead, liability rests with the guardian on the grounds that he/she (supposedly) neglected the child, in terms of both *culpa in educando* and *culpa in vigilando*. No longer does the mere fact of having a *de facto* relationship with the perpetrator of damage give rise to liability (as would be the case with a keeper) – on the contrary, there is no presumption of guilt, but only a presumption of liability. This approach greatly benefits the victim, who is not required to prove the guilt of the parties involved.

A tutor may be exempt from liability only if he/she can prove that he/she could not prevent the incident occurring (thus we cannot speak of objective liability).

This kind of liability presupposes the freedom to move and act and seems best suited to regulating the harmful consequences of harmful events caused by robots such as *iCub* (when the project is completed). It is interesting to note that this set of rules relating to liability presupposes a certain level of material ability in

[180] Article 2048 Italian Civil Code: "Tutors and those who teach a craft or art are responsible for damage caused by unlawful acts of their students and apprentices in the time they are under their supervision. The persons referred to in the preceding paragraphs shall be released from liability only if they prove not to have been able to prevent the fact." (Or in the original Italian: "Il padre e la madre, o il tutore, sono responsabili del danno cagionato dal fatto illecito dei figli minori non emancipati (314 e seguenti, 301, 390 e seguenti) o delle persone soggette alla tutela (343 e seguenti, 414 e seguenti), che abitano con essi. La stessa disposizione si applica all'affiliante. I precettori e coloro che insegnano un mestiere o un'arte sono responsabili del danno cagionato dal fatto illecito dei loro allievi e apprendisti (2130 e seguenti) nel tempo in cui sono sotto la loro vigilanza. Le persone indicate dai commi precedenti sono liberate dalla responsabilità soltanto se provano di non avere potuto impedire il fatto.").

the agent (minor or learning robot), and, thus, of legal subjectivity. Anyone who thinks that legal reasoning is rushing too far ahead must consider another example of cutting-edge robots.

Nao as an ethical being

Nao (developed by the French company Aldebaran Robotics[181]) is a humanoid robot created for the purpose of carrying out functions of assistance. Apart from its skill of being able to communicate with its owner, who may thus teach it new behaviour, and its participation in the RoboCup[182], *Nao* is a very special robot since it is the first one into which an ethical code has been inserted. Its designers have inserted, in an automatic learning algorithm, a series of situations which present ethical problems and their correct solution. Actions are classified on the basis of three principles: beneficence, non-maleficence, and fairness. On the basis of ethical choices preloaded by the programmer, which show how it must act in a certain standard situation, the robot obtains a general rule of ethical conduct. The robot is thus able to independently assess the situation based on this new scale of values and to therefore make the right decision. The robot's scope of use is hospital care. The robot will deal with patients, starting from the standard three situations:

1. Reminding a patient to take his/her drugs but not interfering with his/her refusal to do so unless such a refusal could lead to serious consequences for the patient's health;
2. Deciding who will use the TV remote control, and
3. Delivering food to a patient.

If these three situations all occur at the same time, *Nao* can also make an independent assessment of priority, always based on the three ethical principles above (Anderson and Anderson, 2010).

The existence of *Nao* opens up new scenarios of interplay between robotics, ethics and the law. If a code of ethics can be installed in robots, then it should be a specific responsibility of manufacturers to install that code of ethics in machines which are capable of taking independent decisions. If a manufacturer does not do so, it cannot prove that it could not have avoided a harmful event occurring. A user may also be required to install a code of ethics and conduct, if a robot is equipped with a program which may be managed by a user. Failure to install such codes may be equated with failure to supervise. In any case, these kinds of liability being understood, it must be recognised that producers, owners and users are increasingly taking on the role of external controller of an entity that has ever more autonomy.

---

[181] See http://www.aldebaran-robotics.com/eng/Nao.php.

[182] Robocup is an international robotics competition founded in 1997. The aim is to develop autonomous soccer robots with the intention of promoting research and education in the field of Artificial Intelligence. The name *RoboCup* is a contraction of the competition's full name, 'Robot Soccer World Cup', but there are many other stages of the competition such as 'Search and Rescue' and 'Robot Dancing'. See http://www.robocup.org/.

Conclusions

It may be said that the traditional categories of Italian law regarding tort are strong enough to cope with any problems arising through the use not only of traditional robots but also of these new forms of Artificial Intelligence which are at various stages of development. Legal situations that arise from the use of different types of robot are very similar and can be handled with the traditional categories of Italian tort law. When the cognitive abilities of robots acquire a certain level of development, the creation of an *ad hoc* system of laws may be required. The lowest level of liability is the least specific and is applicable to all types of robots. It includes manufacturer liability for product defects and the liability of a custodian of the artefact concerned. The liability of those who cause damage by using the robot to perform a dangerous activity is transversal to these two types of liability.

At a second level there is the similarity between robots and animals. This similarity stems from the robots' ability to move freely and the emergence of the first adaptive capacities of development and learning. The final level concerns the extraordinary capacity for learning and problem solving of the more sophisticated robots. They may be considered veritable human 'puppies' of man, increasingly similar to us, which we have to educate and train using principles of ethics and conduct.

Having ruled out direct liability of the robot, the cases analysed thus far are all cases of liability attributed to other subjects. At the current level of development it is pointless to punish robots directly, since they do not have the tools with which to understand and develop their own sense of responsibility. Completely equating a robot to a human being would mean continuing to focus on the old concept of strong Artificial Intelligence, a concept which was abandoned in the 1980s.

Robots, however, are no longer simple mechanical objects. In the not too distant future everyone will have a 'personal robot', just as almost everyone now has a personal computer. Today it is virtually unthinkable to leave the house without your mobile phone, or travel without GPS. Such devices are simple but they will continue to develop exponentially. Unbeknown to us, robots are entering all areas of our lives. It is not inconceivable that, sooner or later, they will be given a minimum of subjectivity and *ad hoc* legal status. The degree of legal liability to be attributed to robots directly depends on the level of legal subjectivity they may be given. Numerous issues will arise – from an ethical point of view, what are artificial consciousness and artificial freedom, and when may they be attributed? From a legal standpoint, what would be the most appropriate legal response to a robot being attributed a certain degree of liability? The remedy of autonomous compensation, for example, appears to be rather vague. Another legal issue that arises, and which would be worth studying in depth, concerns the specific role of software, not covered by the Directive on product liability. Although this vision may now seem rather futuristic and virtually unattainable, it is highly likely that it will actually come to pass. The relationship between man and machine is becoming increasingly closer, above all in the medical, rehabilitation and care sectors. Knowledge about the human brain and cognitive development are being used to create robots with ever more sophisticated and responsive Artificial Intelligence. At the same time, studies on the development of cognitive robots could be useful in better understanding the functioning of the human brain.

At the present state of the art in robotics civil liability arising from a robot's actions can be asserted

only on the basis of the above legislation. It must, however, be borne in mind that robots which can move, reason and find solutions to problems are being born ... and are growing.

References

Anderson, M., & Anderson, S.L. (2010). Robot be good: A call for ethical autonomous machines. *Scientific American*.

Asaro, P.M. (2007). Robots and responsibility from a legal perspective. Retrieved from http://www.peterasaro.org/.

Asimov, I. (1950). Runaround. In *I, Robot*. New York: Gnome Press.

Bompani, M. (2009). Parole di iCub. *La repubblica*. Retrieved from http://ricerca.repubblica.it/repubblica/archivio/repubblica/2009/10/30/parola-di-cub.html.

Carreyrou, J. (2010). Surgical robot examined in injuries. *Wall Street Journal Online*, retrieved from http://online.wsj.com.

Comporti, M. (2002). Fatti illeciti: Le responsabilità presunte – Artt. 2044-2048. In P. Schlesinger & F.D. Busnelli (Eds.), *Il codice civile: Commentario*. Milano, IT : Giuffrè.

Comporti, M. (2008). Fatti illeciti: Le responsabilità oggettive – Artt 2049-2053. In P. Schlesinger & F.D. Busnelli, (Eds.), *Il codice civile: Commentario*. Milano, IT : Giuffrè.

Diurni, A. *et al*. (2008). Artt. 2043-2053: fatti illeciti. In P. Cendon, (Ed.), *Commentario al Codice Civile*, Milano, IT: Giuffrè.

Floridi, L. (1999). *Philosophy and computing: An introduction*. London/New York: Routledge.

Kormushev, P., Calinon, S., Saegusa, R., & Metta, G. (2010). Learning the skill of archery by a humanoid robot iCub. Retrieved from http://kormushev.com/papers/Kormushev_Humanoids-2010.pdf.

Invernizzi, G. (2010). Infallibile chirurgo. *Robotica Magazine*, *1*, 28-29.

Ioriatti, E. (2005). Codice civile europeo: Un'occasione per l'Europa sociale. Retreived from http://www.riviste.provincia.tn.it/ppw/Europa.nsf/0/BCC3A6F747B03D36C12570A80036D12F/$FILE/4+codice+civile.pdf?OpenElement.

Marino, D., & Tamburrini, G. (2006). Learning robots and human responsibility. *International Review of Information Ethics*, *6*, 46.

McCarthy, J. (2007). What is artificial intelligence?. Stanford University. Retrieved from

http://www.formal.stanford.edu/jmc/whatisai/.

Oldani, R. (2010a). iCub, il robot bambino. Robotica Magazine, no.1, 5-11.

Oldani, R. (2010b). Arrivano i robonauti. Robotica Magazine, no.1, 20-27.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, New Series*, *59*(236), 433-460.

Veruggio, G. (2007). EURON ROBOETHICS ROADMAP. Plenary Session, CEPE 2007, Seventh International Computer Ethics Conference, July 2007, University of San Diego, USA. Retrieved from http://www.roboethics.org/index_file/Roboethics%20Roadmap%20Rel.1.2.pdf.

Veruggio, G. (2007). The birth of Roboethics. *Leadership medica*, *10*, 6-19. Retrieved from http://www.leadershipmedica.com/sommari/2007/numero_10/Veruggio/Verruggio.pdf.

Von Bar, C. (2009). *Principles of European law: Non-contractual liability arising out of damage caused to another* (Study Group on a European Civil Code). Sellier European Law Publishers.

Chapter 19
Techno-elicitation: Regulating behaviour through the design of robots

Bibi van den Berg
Tilburg University
Tilburg Institute for Law, Technology and Society (TILT)
✉ bibi.vandenberg@tilburguniversity.edu

Abstract In the field of Law & Technology, scholars investigate the legal and regulatory consequences of the advent of new technologies, for example with respect to ICTs, biotechnologies, nanotechnologies, or neurotechnologies. It is important to investigate whether technological developments in these fields require adjustments in existing legal frameworks, and whether technological developments themselves need to be regulated. Moreover, in Law & Technology scholars also investigate the ways in which technological artefacts can be used *to regulate*. This is called 'techno-regulation'.

This paper has two goals. First, I will analyse the concept of techno-regulation and propose that it needs to be broadened. Techno-regulation focuses on the *intentional* influencing of human behaviour through the implementation of values, norms and rules into technological artefacts. However, extensive research in various disciplines has revealed that the design (shape, form, functionality) of technological artefacts greatly affects users' tacit and implicit responses to these artefacts. Since this has direct relevance to the theme of regulation, I propose to widen the reach of techno-regulation by speaking of 'techno-elicitation' instead.

In the second part of this paper, I focus my discussion of techno-regulation and techno-elicitation on the design of robots, which is relatively uncharted territory in the field of Law & Technology.

Keywords robots, techno-regulation, techno-elicitation, social responses, philosophy of design

Introduction

In the previous decades Law & Technology has become an established domain of legal scholarship. This field builds on the realisation that the advent and proliferation of new technologies has an impact on existing legal systems, and affects central (regulatory) values in societies. Hence, technological developments require a response from regulators and legal scholars. In order to find out precisely what response is needed – which of course varies from one technology to the next, and from one institutional, legal and economic system to the next – Law & Technology asks questions such as: What is the impact of technological developments on existing forms of regulation and (bodies of) law? Should the development of new technologies, for example information and communication technologies (ICTs),

biotechnologies, or nanotechnologies, be regulated, and if so, in which ways, or through which means?

The field of Law & Technology has two main areas of focus: the regulation *of* technologies, and regulation *through* technologies. I will discuss these in turn.

Regulation *of* technologies

The majority of research in Law & Technology focuses on the question of whether new technologies require changes to existing legal frameworks, and/or whether the development and proliferation of these new technologies raises new legal problems. Each new technology raises new sets of behaviours, new risks, and new practices of use, and hence legal scholars and governing bodies must investigate whether the use or application of such technologies has consequences that may fall outside existing legal frameworks. The scope of this area of research is vast. To give a few examples, it ranges from studying the effects of the use of information and communication technologies (ICTs) on citizens' privacy, to studying the validity and reach of intellectual property law and patent law in light of the advent of biotechnologies, to investigating the legal consequences of applying neurotechnologies and technologies for human enhancement in various social domains. In all cases researchers focusing on the regulation of technologies ask the following questions:

1. What are the effects, risks, opportunities and dangers resulting from the advent of new technologies, both in direct and in more indirect (or implicit) senses?
2. In which ways, and to which degrees, do existing legal frameworks provide sufficient protection against the possible problems, risks and dangers that may arise in the slipstream of these developments?
3. If legal frameworks are found to provide insufficient protection in one or more areas, then how can these frameworks be adjusted, so as to solve the problem?
4. And finally, especially in the case of technological developments that are considered inherently dangerous or risky, should the development of specific technologies as such be regulated, or the institutional or organisational environment into which they will enter, so as to ensure as safe an application as possible?

Asking and answering these questions, it is important to note, is always, and principally, a *contextual* enterprise. As Bert-Jaap Koops writes:

> Questions of technology regulation always have to take into account the location both of the technology and regulatory attempts, so that relevant socio-cultural, legal, economic, and institutional factors associated with that place can be factored in. (Koops, 2008, p. 314)

Regulation *through* technologies

As said, the majority of scholars in the field of Law & Technology study questions surrounding the regulation *of* technologies. Increasingly, however, a second domain of focus is gaining prominence: that of regulation *through* technologies. Lawrence Lessig has famously argued that technologies can also be used *to regulate*, i.e. to steer and guide the behaviour of individuals (Lessig, 2006). This has come to be known as 'design-based regulation' (Brownsword & Yeung, 2008) or 'techno-regulation' (Brownsword, 2008; Leenes, 2010). Techno-regulation studies the ways in which technologies can be used as regulatory tools (Brownsword & Yeung, 2008), i.e. as a means to influence the behaviours of individuals by implementing regulatory values, norms and standards into technological devices (Koops, 2008). Note that for scholars in Law & Technology 'regulation' relates to the *intentional* influencing of human behaviour. This means that techno-regulation, to them, revolves around the ways in which regulators – be they governments or industry or any other party – may attempt to evoke behaviours in regulatees through the intentional implementation of norms and standards into technological artefacts. Below I will question this exclusive focus on intentional influencing. For now, however, let's look at some examples of techno-regulation to shed light on its meaning and role in various social contexts.

One of the most oft-cited examples of techno-regulation is that of the use of speed bumps in traffic (Brownsword, 2008; Latour, 1992; Leenes, 2010; Yeung, 2008). Speed bumps are only one means of ensuring that drivers will adhere to a designated maximum speed in a certain area. Regulators can also choose to use traffic signs to the same end. However, the use of a speed bump regulates the driver's speed in a much more direct, and binding, way: a speed bump leaves much less room for being 'disobedient' than using traffic signs. After all, driving over a speed bump at high speed is physically uncomfortable and may damage the driver's car. Driving past a traffic sign at high speed does not affect the driver directly in this way. Hence, when using a speed bump chances are that drivers will be much more inclined to adhere to the traffic rules than when using a traffic sign. *By* design and *through* design speed bumps encourage drivers to stay within the speed limits set by a regulator.

Another example of techno-regulation is that of the use of DVD region codes. DVDs, Leenes writes "*generally contain various mechanisms of Digital Rights Management, which*

define what a user can and cannot do with the DVD"[183] (Leenes, 2010, p. 11). Media industries have divided the globe into nine different regions, so that DVDs can be marketed with different content, for different prices, and with different release dates in each region (Leenes, 2010). DVDs that work in one region, say Europe (region 2), will not play on DVD players in another, say the US (region 1), and vice versa. This is a clear example of regulation *through* technology – the software in the machine, and the code on the disc, jointly ensure that viewers can only watch those DVDs they are 'allowed' to watch, according to the industry's regulatory plans. Leenes writes: "*The technology enforces adherence to the rules by means of the software that is implemented into the machine. The enforcement is (almost) perfect*." (Leenes, 2010, p. 11)

These two examples show that techno-regulation focuses on implementing rules, values, norms and standards *into the architecture*, or *code* in the case of software, of the artefact itself, thus ensuring that obeiance to laws and regulations is obtained. Morgan and Yeung write: "*code-based (or architecture-based) techniques* [seek] *to eliminate undesirable behaviour by designing out the possibility for its occurrence*" (Morgan & Yeung, 2007, p. 102). Or in the words of Brownsword:

> …techno-regulation […] functions in such a way that regulatees have no choice at all but to act in accordance with the desired regulatory pattern – it is the difference, for example, between systems that make it physically impossible to exit the Underground (or Metro) without a valid ticket and low level barriers that make it more difficult (but not impossible) to do so… (Roger Brownsword, cited in Morgan & Yeung, 2007, p. 103)

Note that not just the specific form of regulation implemented into a technological artefact, but also the level of *regulability as such* is a design choice: "*Different code makes differently regulable* [technologies]. *Regulability is thus a function of design*." (Lessig, 2006, p. 34)

Techno-elicitation: Widening the reach of Law & Technology

In the previous section I argued that scholars in the field of techno-regulation focus primarily on the *intentional* influencing of human behaviour through the design of technologies. This applies, first and foremost, to those investigating the ways in which

[183] Translated by the author.

technologies can/ought to be regulated, but also to those focusing on techno-regulation.[184] In itself this is not surprising. After all, lawyers and regulators seek to find ways to explicitly channel behaviour, to keep it within the boundaries of the law. Therefore, 'regulation', to legal scholars, means "*the intentional influencing of someone's or something's behaviour*" (Koops, 2008). What this entails, however, is that *un*intentional forms of influencing, which may arise for example as a side-effect in the design of technologies, or forms of influencing that may steer individuals in more implicit ways, largely fall outside the scope of (techno-)regulation research.

To my mind, this omission is unfortunate, and in this paper I will explain why this is so. I argue that it would be good to increase the scope of research on techno-regulation beyond intentional influencing alone, because human behaviour is often strongly shaped, steered and affected in more subtle, implicit, and even unconscious ways by technological artefacts as well. Over the past decades a significant corpus of research in different disciplines, including engineering, computer science, human-computer interaction (HCI), human-robot interaction (HRI), science and technology studies (SST), and philosophy of technology, has consistently shown just how ubiquitous and important the *unintended*, *implicit* and *automatic* elicitation of human behaviours is in relation to technological artefacts. Technologies have been shown to have 'persuasive powers' (Fogg, 2003), which sometimes may be designed into them explicitly, but sometimes also operate in more subtle ways. Moreover, technologies contain 'scripts' (Akrich, 1992; Gjøen & Hård, 2002; MacKenzie & Wajcman, 1999; Oudshoorn & Pinch, 2003; Oudshoorn, Rommes, & Stienstra, 2004; Van den Berg, 2008, 2010), which delineate their use space, and invite certain types of behaviour, while constraining others (Hildebrandt, 2008a, 2008b; Latour, 1992; Winner, 1980). Or in different terms, technologies 'afford' certain actions, and restrict other behaviours, and hence implictly shape the behaviours of users (Gaver, 1991, 1996; Gibson, 1986; McGrenere & Ho, 2000).

What's more, research has also shown that human beings have strong tendencies to 'anthropomorphise' technologies (Bartneck, Kulic, Croft, & Zoghbi, 2009; Duffy, 2003; Nass, Steuer, Tauber, & Reeder, 1993; Turkle, 1984), to ascribe intentions and agency to these inanimate objects. This applies even to quite 'simple' artefacts, which do not display complex

[184] While legal scholars writing on techno-regulation often acknowledge explicitly that technological artefacts may also *unintentionally*, *subtly*, and *implicitly* regulate human behaviour as well (see for example Brownsword, 2008; Leenes, 2010; Yeung, 2008), their work focuses on the *intentional* influencing of human behaviour through design.

or very varied patterns of behaviour. One of the most famous examples to show how easy it is to invoke a tendency to anthropomorphise in humans is Joseph Weizenbaum's computer program ELIZA, which mimicked the behaviours of a Rogerian psychoanalyst (Weizenbaum, 1966). ELIZA consisted of a simple textual interface, through which individuals could 'converse' with this virtual therapist. The program used a limited set of conversion rules to turn users' phrases into questions, thus invoking the idea that the 'therapist' followed up on whatever they shared with a next question. Weizenbaum was shocked to find out how convincing his program turned out to be, i.e. how strongly users anthropomorphised this simple software program. He said:

> I was startled to see how quickly and very deeply people conversing with [ELIZA] became emotionally involved with the computer and how unequivocally they anthropomorphized it. Once my secretary, who had watched me work on the program for many months and therefore surely knew it to be merely a computer program, started conversing with it. After only a few interchanges with it she asked me to leave the room. Another time, I suggested I might rig the system so that I could examine all the conversations anyone had had with it, say, overnight. I was promptly bombarded with accusations that what I proposed amounted to spying on people's most intimate thoughts; clear evidence that people were conversing with the computer as if it were a person who could be appropriately and usefully addressed in intimate terms. (Joseph Weizenbaum, quoted in Kerr, 2004, p. 305)

Note that it is not just computer technologies that easily evoke anthropomorphisation. Philosopher of technology Don Ihde reminds us that at times we also tend to 'animate' cars, almost approaching them as if they are a kind of 'spirited horse', and that we 'compete' with virtual characters in video games as if they were real others (Ihde, 1990; also see Verbeek, 2005).

Yet another branch of research has shown that, at times, we even respond to technological artefacts in *social* and *emotional* ways (Breazeal, 2002; Dautenhahn, 2007; Dautenhahn, Bond, Canamero, & Edmonds, 2002; Picard, 1997; Turkle, 2007). This has led to a number of research initiatives investigating what exactly triggers such social or emotional responses to machines in humans – not only to, for example robots, but also to computers and televisions. Quite contrary to what one might expect Reeves and Nass' extensive research in this domain consistenly reveals that humans, in fact, need only very minimal cues to invoke them. Even machines that do not even remotely look human (e.g., ordinary desktop computers), or display complicated behaviours (e.g., relatively simple software programs) evoke basic social mechanisms, such as a sense of politeness or of teamwork in users (Reeves & Nass, 1996).

Over the years, many explanations have been given for all of these implicit human responses to technological artefacts. Most often, these tendencies are explained by referring to our species' evolutionary 'social hardwiring': because we are social, emotional beings

through and through, we automatically use our repertoire of social and emotional responses in our interactions with technological artefacts (Nass & Moon, 2000; Nass, Steuer, & Tauber, 1994; Picard, 1997; Reeves & Nass, 1996).

What this vast body of research from various disciplines consistently shows, then, is that through their design technological artefacts may influence the behaviours of human beings in a variety of subtle, and implicit ways. This is relevant to those interested in techno-regulation as well. While users may sometimes be aware of technologies' powers of influence [read: regulatory powers], and may consciously accept or reject such regulation, apparently humans' behaviours can also be influenced [read: regulated] in more implicit and tacit ways. Perhaps, then, the scope of research on techno-regulation so far has been too narrow and ought to be widened, to include both intentional influencing and more tacit forms thereof. I propose to do just that, by replacing the notion of 'techno-regulation' with what I call 'techno-elicitation'. Techno-elicitation relates to *all forms of evoking human behaviour through technological design*. It is a scale of responses in users, running from explicit and conscious ones to implicit, and tacit evocations.

Users and designers

So far, in this article we've focused on the role technologies may play in either intentionally or implicitly influencing users. Techno-elicitation covers the entire range of behaviours users may display in response to (influences of) technological artefacts. However, studies have also shown that it is not just *users*' responses to the affording and constraining powers of technologies that are often implicit and tacit. Research in Science & Technology Studies (Akrich, 1995; Oudshoorn & Pinch, 2003), Actor Network Theory (Latour, 1992, 2005; Latour & Venn, 2002), value-sensitive design (Friedman, 1997; Friedman & Kahn Jr., 2006; Friedman, Kahn Jr., & Borning, 2002), and philosophy of design (Kroes, Light, Vermaas, & Moore, 2009; Verbeek, 2005) consistently reveals that *designers*, too, are often unaware of values, norms and stereotypes they embed into the artefacts they create. In many cases designers use implicit user models in the design process. Van Oost illustrated this in research on the values embedded into male and female shavers, which tacitly reflect ideas on gender differences: male shavers are grey and black, contain dials and screws, can be opened up and taken apart. Female shavers, in contrast, come in pastel colours, have smooth and curvy shapes, lack dials and switches, and cannot be taken apart (Van Oost, 2003). These differences are based on tacit assumptions on the part of he designers, Van Oost says, and they reflect stereotypical ideas on gender and technology use: men like technologies, and therefore want a shaver that looks as 'technological' as possible, whereas women are afraid of technology, and hence prefer shavers that look more like a cosmetics product than a technological artefact. Van Oost concludes:

> ...the gender script of the [female shaver] inhibits [...] the ability of women to see

> themselves as interested in technology and as technologically competent, whereas the gender script of the [males shavers] invites men to see themselves that way. In other words: Philips [, the manufacturer,] not only produces shavers but also gender. (Van Oost, 2003p. 207)

One of the key findings in Van Oost's research was that the designers themselves were not aware of the fact that they had embedded stereotypical values into their design. One explanation why such value-embedding may easily be tacit and implicit in designers is what Oudshoorn has called 'I-methodology' (Oudshoorn & Pinch, 2003), i.e. designers' tendency to take themselves, their own needs, attitudes, preferences and capacities, as the main point of reference in design (Van den Berg, 2010).

What this reveals is that the concept of techno-elicitation, as we've defined it so far – focusing only on the user side – is still too narrow. Techno-elicitation, we must conclude, is a spectrum running from *intentional* and *explicit* evocation on one end (techno-regulation), to *implicit*, *accidental* and *unintentional* elicitation on the other (scripts, animism etc.), and it holds for both the *users* and the *designers* of technological artefacts. To complicate things further, different technologies all have their own medium-specific characteristics, which means that different technologies lead to different forms of techno-elicitation. In order to shed light on the workings of techno-elictation we need to investigate its occurrences and effects in different technological domains, then. In the second part of this article I will attempt to do so by focusing on regulation and robotics.

Regulating robotics

As we saw at the beginning of this article, technological developments require scrutiny on the part of legal scholars, to investigate whether laws and regulations need adjustment, to determine whether their design and/or proliferation needs to be regulated, and to come to an understanding of the regulatory powers of these technologies. Against this background, legal scholars have also turned to regulatory questions surrounding the advent of (increasingly) autonomous technologies, robotics and artificially intelligent machines. In fact, they were surprisingly early to realise that the creation of such intelligent, autonomously operating artefacts needed to be evaluated critically from a legal point of view as well. The earliest articles written in this field date from the beginning of the 1980s – a time when the realisation of artificially intelligent machines was a distinctly more remote possibility than it is today. Since that time, a serious body of literature has been created on the legal issues that may arise in a world inhabited by robots (as well as people).

In this body of literature, legal scholars have largely focused on three key themes: liability, the legal status of robots, and rights for robots. First of all, the advent of robotic and autonomous technologies raises questions regarding *liability* when things go wrong: who is responsible for a robot's behaviours? Do robots fall under product liability, and hence can we

hold manufacturers responsible for the damage they may cause? Or should robots be considered a special type of products, for whose behaviours producers cannot be held responsible, because, for example, their machinery is so complex that their behaviours will be inherently unpredictable? Or because neural networks enable them to learn new things that nobody has programmed into them? Or because so many companies, individuals and groups contribute to the creation of these machines that it becomes impossible to hold one company, individual or group responsible for their behaviours (Wallach & Allen, 2009)?[185] One solution that legal scholars propose to keep responsibility in the hands of humans while acknowledging some sense of 'agency' in robots, is to use legal constructions such as those pertaining to parents and children, owners and their wild animals, principles and agents in commerce, or employers and employees, and apply these to liability issues surrounding robots. In this way, the owners of robots would be held responsible for any damage these machines may do (Lehman-Wilzig, 1981). What complicates the study of liability and robotics is that issues of liability vary greatly across domains of application: robotic cars may have different legal provisions (i.e. in traffic law) than robots for the household (i.e. consumer law), and those used in the warfare (i.e. international law). Moreover, laws on liability vary from country to country, which further complicates the study of liability issues in the domain of robotics.[186]

A second domain of study in law and robotics relates to the question of the *legal status* of robots and other intelligent and/or autonomous machines. The central question here is: should robots be given a legal status, other than being a mere object, and hence become 'legal persons', and if so, what are the requirements they should meet in order to be granted such a status? Granting robots (or any other nonhumans) with legal status, and calling them a legal person, may seem counter-intuitive to non-lawyers at first, but in fact, several authors point out that legal personhood certainly isn't reserved for humans only (Calverley, 2008; Koops, Hildebrandt, & Jaquet-Chiffelle, 2009; Solum, 1992). Koops, Hildebrandt and Jacquet-Chiffelle write: "*In most modern legal systems, legal personhood is attributed to associations, funds or even ships*" (Koops et al., 2009, p. 9), and companies, trusts and other collectives are also recognised as legal persons by most legal systems. All of these (nonhuman) entities are treated as separate, autonomous entities by the law, rather than as an aggregate of the people that make up these entities, or as a collection of people behind them (Calverley, 2008;

[185] Also see Wendell Wallach's article in this volume.

[186] Chiara Boscarato's article in this volume discusses liability and robotics under Italian law.

Solum, 1992). Moreover, who or what counts as a legal person turns out to be a rather changeable, fluid category when viewed from a historical perspective. For centuries all sorts of nonhumans have played a role in Western law, from which we have recently eliminated them. For instance, there is a long series of animal species that have been tried in court throughout history, ranging from donkeys and beetles to rats, grasshoppers, dolphins and eels (Teubner, 2006). In a famous case the rats were exonerated on the grounds that it was impossible to set a date for their appearance before the judge (Teubner, 2006). Certain buildings, such as Roman temples and Medieval churches also used to have legal rights in various cultures of the past (Solum, 1992). And it is not just animals and structures that have figured in legal cases throughout history – so have all sorts of ghosts and gods, and a wide variety of other visible and invisible 'influences' (allegedly) affecting everyday life. More importantly, we also need to consider the fact that a significant portion of *human beings* today have rights that up until very recent times did not. Think for instance of women (Magnani, 2007), slaves (Lehman-Wilzig, 1981), children, foreigners and refugees, or people with disabilities or mental illnesses. These examples show that the category of legal personhood is not set in stone. At different times, different entities have been considered as legal persons or not. According to legal scholars, this means that we ought to at least consider the question of applying the term 'legal person' to robots, and to autonomously operating or artificially intelligent machines as well.

A third theme in research on robotics and regulation revolves around the question of *legal rights for robots* (Teubner, 2006). The debate in this area mainly focuses on comparisons between humans, as full bearers of rights, animals, as bearers of some rights in certain jurisdictions, and machines, which up until this point in time do not have rights. Deciding whether or not to grant such rights, Solum argues, would depend on both the rights themselves (e.g., the right to freedom of expression or the right to emancipation) and on the justification used for granting that right (Solum, 1992).

One line of reasoning for *withholding* all constitutional rights from autonomous, smart technologies without further justification is to claim that such rights can be given only to humans, full stop. Solum calls this the 'anthropocentric argument', which comes down to saying "*We are humans. Even if* [artificially intelligent machines] *have all the qualities that make us moral persons, we shouldn't allow them the rights of constitutional personhood because it isn't in our interest to do so*" (Solum, 1992, p. 1260). Although this may sound intuitive and express deeply held feelings by many, Solum rightly points out that this is a very shady moral argument, "*akin to American slave owners saying that slaves could not have constitutional rights simply because they were not white or simply because it was not in the interests of whites to give them rights*" (Solum, 1992, p. 1261). An even more dubious version of this argument is the 'paranoid anthropocentric argument', which claims that we should not give these nonhumans rights because they might become so powerful they would take over the world. This is an argument we should not take seriously at all, says Solum, because

> …the danger seems remote, but if the danger were real it would not be an argument against granting [artificially intelligent machines] legal personhood. If [these machines] really will pose a danger to humans, the solution is not to create them in the first place. (Solum, 1992, p. 1261)

It appears, then, that at least in theory we cannot rule out that robots and other artificially intelligent machines may one day acquire legal status and be given legal rights in some form or other, that is, if they meet the requirements placed on humans and some nonhumans to qualify for these matters.

Techno-regulation and robots: Uncharted territory

The reader may have noticed that all three of the research themes discussed above fall within the domain of 'regulation *of* technologies' that I discussed at the beginning of this article. They all focus on the question of how advances in robotics fit within existing regulatory frameworks and bodies of law, and whether changes are required in those frameworks and bodies of law to meet the new social and legal demands created by the advent of such technologies. Alternatively, they focus on questions regarding the need (or lack thereof) or regulating the development and deployment of robotics technologies.

Why would it be relevant to study questions of techno-regulation and techno-elicitation in relation to robotics in the first place? I will answer this question by discussing two domains of application in robotics: healthcare and the military.

Robots in healthcare

A recent OECD report on healthcare spending stated that "*in all OECD countries total spending on healthcare is rising faster than economic growth*." (OECD, 2010). The World Health Organization (WHO) warns that while life expectancy is increasing, simultaneously birth rates are decreasing in most countries (WHO, 2010). This challenges existing healthcare systems: more people need healthcare services, yet fewer humans are available to provide those services.

One area of research and business rapidly developing to face this challenge is that of *healthcare robotics*. Healthcare robots, or 'carebots', could conduct various care tasks, such as delivering medication and food, monitoring, lifting or transporting patients, and providing companionship. Healthcare robots can also be used for therapeutic ends. Interaction with

robotic pets, such as Sony's AIBO[187] or the robot seal Paro[188], has been empirically shown to have a positive effect on the activity and social interaction levels in elderly people, to improve patients' moods, and to reduce stress levels and loneliness (Banks, Willoughby, & Banks, 2008; Broekens, Heerink, & Rosendal, 2009; Stiehl et al., 2005; Wada & Shibata, 2008; Wada, Shibata, Mushi, & Kimura, 2008).

Applying robots in care practices for the elderly and the sick also has a wide range of ethical consequences. In recent years a number of studies have been conducted on the ethical aspects of the application of robots in healthcare situations (Borenstein & Pearson, 2010; Coeckelbergh, 2009; Tiwari, Warren, & Day, 2010)[189]. These focus on, for instance, qualitative differences between care provided by humans and by robots, on the way the central values of our healthcare system, and our ideas on care, are affected by the application of healthcare robots, and on the requirements – both social, practical, emotional and ethical – that robots must meet if we are to allow them to care for our elderly and sick.

Yet studying the ethical aspects of applying robots to healthcare situations alone is not enough. Precisely because socially and emotionally complex contexts in which healthcare robots must operate, caring for patients in vulnerable situations, we must also elucidate the ways in which the design of healthcare robots, in terms of their physical form and functionalities, has a bearing on the behavioural responses they may elicit. As we have seen in this article, such behavioural responses may be evoked explicitly and intentionally, but also more implicitly and perhaps at times even unintentionally on the part of the designer. Moreover, users may be explicitly aware of the fact that certain behaviours are invoked by (the design of) healthcare and other robots, yet they may also be so subtle that they escape users' awareness.

Investigating the consequences of explicit (regulatory) design choices with respect to these machines is important for two reasons. First, it increases our ability to develop robots that uphold central values in healthcare practices, such as respecting patients' autonomy, privacy and integrity. Second, it contributes to defining the role, meaning and ethical 'bearing' of healthcare robots. Since technologies "*are by definition value-laden systems and designing such systems is, by definition, a value-laden activity*" (Kroes et al., 2009, p. 13), explicating (regulatory) design choices can contribute to designing legally, socially and ethically sound

[187] See http://support.sony-europe.com/aibo/index.asp

[188] See http://www.parorobots.com/.

[189] Also see Aimee Van Wynsberghe's article in this volume.

healthcare robots.

Robots in warfare

Research and development of robots for military purposes – both surveillance and warfare – has sped up and expanded more than any other area of in robotics in recent times. A significant number of robots is currently participating in the war in Afghanistan, in a variety of roles, ranging from finding explosives to patrolling the skies. While human beings are still always 'in the loop' when it comes to making final decisions in combat and in surveillance today, several researchers suggest that we are rapidly moving towards an era in which robot soldiers will engage in combat autonomously (Arkin, 2009; Krishnan, 2009; Singer, 2009). The fact that there is a wide range of thorny ethical and legal issues to be addressed has not gone unnoticed to these authors and others.[190] Debates run high regarding the question of a need for, and possibility of, implementing morality into robots[191] that participate in warfare, to turn them into 'ethical warriors', and of course, questions of liability, of international law (*jus in bello*), and of 'just wars' are on the agenda as well.

Many authors discuss the design and functionality that robot soldiers ought to have. What they implicitly say is that the design of these machines, the code we implement into them, has far-reaching consequences for the output, the behaviours they will generate in the real world. And now is the time to think about these matters: as developments in the creation of such machines are picking up speed. Or in the words of Lessig:

> Choices among values, choices about regulation, [and] about control […] – all this is the stuff of politics. Code codifies politics, and yet, oddly, most people speak of code as if it were just a question of engineering. Or as if code is best left to the market. Or best left unaddressed by government. […] How the code regulates, who the code writers are, and who controls the code writers – these are questions on which any practice of justice must focus… (Lessig, 2006, p. 78-79)

As with healthcare robots, here, too, the central aim is to generate discussion on the values we embed into machines, and the effects this may have in the settings in which they will be deployed. And here, too, studying the ethical aspects of applying robots to war is not

[190] Also see Andreas Matthias' article in this volume.

[191] For more on the foundations of building morality into machines, see the articles of Samir Chopra, Steve Torrance, and David Jablonka in this volume.

enough. Since military robots operate alongside human beings, the same types of implicit behavioural responses that have been discussed throughout this article also appear in soldiers in response to their interactions with such robots. Singer describes a clear case in point: during the war in Iraq, soldiers who had operated alongside a PackBot to find and dismantle explosives, became strongly emotionally attached to this machines, and were deeply saddened when one bad day it was blown up by a roadside bomb. They had named the robot Scooby-Doo, and had gone through so many difficult missions with it, in which it saved their lives a number of times. The soldiers were "*very upset*" when they learnt that it could not be repaired (Singer, 2009, p. 338). Singer writes:

> …while new technologies are breaking down the traditional soldierly bonds, entirely new bonds are being created in unmanned wars. People, including the most hardened soldiers, are projecting all sorts of thoughts, feelings, and emotions onto their new machines, creating a whole new side to the experience of war. […] Soldiers […] are truly bonding with these machines. (Singer, 2009, p. 338)

This example shows that in the case of military robots, too, it is important to get a better understanding of the ways in which the design of such robots may influence the behaviours of the individuals that have to work with it in the field. Both the functionality and physical shape of these machines must be taken into account to get a clearer grasp on the forms of techno-elicitation they invoke.

The same two reasons why it is important to investigate techno-elicitation in healthcare robots also apply to military robots, then. First, a better understanding of the workings of techno-elicitation in soldiers and other military personnel increases our ability to design and develop machines that meet their (all-to-human) social and emotional needs, and respect values such as comradery and teamwork in the army. Second, making implicit behavioural responses explicit, and designing these machines to meet actual needs and preferences, will lead to more socially and ethically attuned military robots.

Conclusion

In this article I set out to investigate some boundaries of the concept of 'techno-regulation', which is one of the key focal areas in Law & Technology. Techno-regulation focuses on the ways in which technologies can be used as regulatory tools, as instruments to intentionally steer and influence the behaviours of individuals. While I firmly believe in the enterprise of techno-regulation research as such, I have argued there is a need to widen the reach of this field of study, by also including implicit, tacit forms of influencing people's behaviours through technologies, and, moreover, to not only focus on the regulatory responses invoked in users, but also on the ways in which designers (sometimes intentionally, but often also tacitly) implement values, stereotypes and norms into

technologies. I have shown that developing a clearer conceptual understanding of the full range of techno-elicitation leads to a better grasp of techno-regulation as one of its manifestations, and this, in the process, consolidates the academic enterprise of Law & Technology.

In the second part of this chapter I have discussed regulation and robotics. After a discussion of the current legal debates in this field, I have used the design of robots in two different domains – healthcare and the military – as an empirical lens to contribute to a deeper understanding of the concept of 'techno-regulation' and other forms of behaviour elicitation. I conclude that a deeper understanding of the explicit and implicit regulatory powers of robots in these domains may contribute to more ethically, socially and legally sounds design of these machines.

References

Akrich, M. (1992). The de-scription of technical objects. In W.E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 205-224). Cambridge (MA): MIT Press.

Akrich, M. (1995). User representations: Practices, methods and sociology. In A. Rip, T.J. Misa, & J. Schot (Eds.), *Managing technology in society: The approach of constructive technology assessment* (pp. 167-184). London; New York (NY): Pinter Publishers.

Arkin, R.C. (2009). *Governing lethal behavior in autonomous robots*. Boca Raton: CRC Press.

Banks, M.R., Willoughby, L.M., & Banks, W.A. (2008). Animal-assisted therapy and loneliness in nursing homes: Use of robotic versus living dogs. *Journal of the American Medical Directors Association, 9*(3), 173-178.

Bartneck, C., Kulic, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics, 1*(1), 71-81.

Borenstein, J., & Pearson, Y. (2010). Robot caregivers: Harbingers of freedom for all? *Ethics and Information Technology, 12*(3), 277-288.

Breazeal, C.L. (2002). *Designing sociable robots*. Cambridge, Mass.: MIT Press.

Broekens, J., Heerink, M., & Rosendal, H. (2009). The effectiveness of assistive social robots in elderly care: A review. *Gerontechnology Journal, 8*(2), 94-103.

Brownsword, R. (2008). So what does the world need now? Reflections on regulating technologies. In R. Brownsword & K. Yeung (Eds.), *Regulating Technologies: Legal*

futures, regulatory frames and technological fixes (pp. 23-49). Oxford: Hart.

Brownsword, R., & Yeung, K. (2008). Regulating technologies: Tools, targets and thematics. In R. Brownsword & K. Yeung (Eds.), *Regulating Technologies: Legal futures, regulatory frames and technological fixes* (pp. 3-23). Oxford: Hart.

Calverley, D.J. (2008). Imagining a non-biological machine as a legal person. *Artificial Intelligence & Society, 22*, 523-537.

Coeckelbergh, M. (2009). Health care, capabilities, and AI assistive technologies. *Ethical Theory and Moral Practice, 13*, 181-190.

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B, 362*(1480), 679-704.

Dautenhahn, K., Bond, A.H., Canamero, L., & Edmonds, B. (Eds.). (2002). *Socially intelligent agents: Creating relationships with computers and robots*. Boston, Mass.: Kluwer Academic Publishers.

Duffy, B.R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42*, 177-190.

Fogg, B.J. (2003). *Persuasive technology: Using computers to change what we think and do*. Amsterdam; Boston: Morgan Kaufmann Publishers.

Friedman, B. (1997). *Human values and the design of computer technology*. Stanford, Calif.; Cambridge; New York: CSLI Publications; Cambridge University Press.

Friedman, B., & Kahn Jr., P.H. (2006). Value sensitive design and information systems. In M.E. Sharpe (Ed.), *Human-Computer Interaction and Management Information Systems: Foundations* (pp. 348-372).

Friedman, B., Kahn Jr., P.H., & Borning, A. (2002). *Value sensitive design: Theory and methods*: University of Washington, Dept. of Computer Science & Engineering.

Gaver, W.W. (1991). *Technology Affordances.* Paper presented at the SIGCHI Conference on human factors in computing systems: Reaching through technology, New Orleans (LA).

Gaver, W.W. (1996). Affordances for interaction: The social is material for design. *Ecological Psychology, 8*(2), 111-129.

Gibson, J.J. (1986). *The ecological approach to visual perception*. Hillsdale (NJ): L. Erlbaum Associates.

Gjøen, H., & Hård, M. (2002). Cultural politics in actions: Developing user scripts in relation to the electric vehicle. *Science, Technology & Human Values, 27*(2), 262-281.

Hildebrandt, M. (2008a). A vision of ambient law. In R. Brownsword & K. Yeung (Eds.), *Regulating Technologies* (pp. 175-191). Oxford: Hart.

Hildebrandt, M. (2008b). Legal and technological normativity: More (and less) than twin sisters. *Technè, 12*(3), 169-183.

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Bloomington (IN): Indiana University Press.

Kerr, I.R. (2004). Bots, babes and the Californication of commerce. *University of Ottawa Law & Technology Journal*, 287-324.

Koops, B.-J. (2008). Criteria for normative technology: The acceptability of 'Code as law' in light of democratic and constitutional values. In R. Brownsword & K. Yeung (Eds.), *Regulating Technologies: Legal futures, regulatory frames and technological fixes* (pp. 157-175). Oxford: Hart.

Koops, B.-J., Hildebrandt, M., & Jaquet-Chiffelle, D.-O. (2009). *Bridging the accountability gap: Rights for new entities in the information society? [Deliverable 17.3]*: FIDIS - Future of Identity in the Information Society.

Krishnan, A. (2009). *Killer robots: Legality and ethicality of autonomous weapons*. Farnham, UK/Burlington, VT: Ashgate.

Kroes, P., Light, A., Vermaas, P.E., & Moore, S.A. (2009). *Philosophy and design: From engineering to architecture*. Dordrecht: Springer Science + Business Media B.V.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W.E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 225-259). Cambridge (MA): MIT Press.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford (UK); New York (NY): Oxford University Press.

Latour, B., & Venn, C. (2002). Morality and technology: The end of the means. *Theory, Culture & Society, 19*(5/6), 247-260.

Leenes, R. (2010). *Harde lessen: Apologie van technologie als reguleringsinstrument*. Tilburg: Universiteit van Tilburg.

Lehman-Wilzig, S.N. (1981). Frankenstein unbound: Towards a legal definition of Artificial Intelligence. *Futures, 13*(6), 442-457.

Lessig, L. (2006). *Code: Version 2.0* ([2nd ed.). New York: Basic Books.

MacKenzie, D.A., & Wajcman, J. (Eds.). (1999). *The social shaping of technology* (2nd ed.). Buckingham (UK); Philadelphia (PA): Open University Press.

Magnani, L. (2007). *Distributed morality and technological artifacts.* Paper presented at the Human Being in Contemporary Philosophy, Volgograd (Russia). Retrieved from http://www.volgograd2007.goldenideashome.com/2%20Papers/Magnani%20Lorenzo%20p.pdf.

McGrenere, J., & Ho, W. (2000). *Affordances: Clarifying and evolving a concept.* Paper presented at the Graphics Interface Conference, Montreal, Quebec (Canada). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.153.8239&rep=rep1&type=pdf.

Morgan, B., & Yeung, K. (2007). *An introduction to law and regulation: Text and materials*. Cambridge, UK/New York: Cambridge University Press.

Nass, C.I., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81-103.

Nass, C.I., Steuer, J., & Tauber, E.R. (1994). *Computers are social actors.* Paper presented at the Computer-Human Interaction (CHI) Conference: Celebrating Interdependence 1994, Boston (MA).

Nass, C.I., Steuer, J., Tauber, E.R., & Reeder, H. (1993). *Anthropomorphism, agency, and ethopoeia: Computers as social actors.* Paper presented at the Computer-Human Interaction (CHI) Conference 1993, Amsterdam (The Netherlands).

OECD. (2010). OECD Health Data 2010.

Oudshoorn, N., & Pinch, T.J. (2003). *How users matter: The co-construction of users and technologies*. Cambridge (MA): MIT Press.

Oudshoorn, N., Rommes, E., & Stienstra, M. (2004). Configuring the user as everybody: Gender and design cultures in information and communication technologies. *Science, Technology & Human Values, 29*(1), 30-63.

Picard, R.W. (1997). *Affective computing*. Cambridge (MA): MIT Press.

Reeves, B., & Nass, C.I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Stanford (CA); New York (NY): CSLI Publications/Cambridge University Press.

Singer, P.W. (2009). *Wired for war: The robotics revolution and conflict in the twenty-first century*. New York: Penguin Press.

Solum, L.B. (1992). Legal personhood for Artificial Intelligences. *North Carolina Law Review, 70*, 1231-1288.

Stiehl, W.D., Lieberman, J., Breazeal, C., Basel, L., Lalla, L., & Wolf, M. (2005). Design of a therapeutic robotic companion for relational, affective touch, *2005 IEEE International Workshop on Robots and Human Interactive Communication* (pp. 408-415): IEEE.

Teubner, G. (2006). Rights of non-humans? Electronic agents and animals as new actors in politics and law. *Journal of Law and Society, 33*(4), 497-521.

Tiwari, P., Warren, J., & Day, K.J. (2010). Some non-technology implications for wider application of robots assisting older people. *Health Care and Informatics Review Online, 14*(1), 2-11.

Turkle, S. (1984). *The second self: Computers and the human spirit*. New York (NY): Simon and Schuster.

Turkle, S. (2007). *Evocative objects: Things we think with*. Cambridge (MA): MIT Press.

Van den Berg, B. (2008). Self, script, and situation: Identity in a world of ICTs. In S. Fischer-Hübner, P. Duquenoy, A. Zuccato & L. Martucci (Eds.), *The future of identity in the information society: Proceedings of the third IFIP WG 9.2, 9.6/11.6, 11.7/FIDIS International Summer School on the Future of Identity in the Information Society* (pp. 63-77). New York, NY, USA: Springer.

Van den Berg, B. (2010). *The situated self: Identity in a world of Ambient Intelligence*. Nijmegen: Wolf Legal Publishers.

Van Oost, E. (2003). Materialized gender: How shavers configure the users' femininity and masculinity. In N. Oudshoorn & T.J. Pinch (Eds.), *How users matter: The co-construction of users and technologies* (pp. 193-209). Cambridge (MA): MIT Press.

Verbeek, P.-P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. University Park (PA): Pennsylvania State University Press.

Wada, K., & Shibata, T. (2008). Social and psychological influences of living with seal robots

in an elderly care house for two months, *6th International Conference of the International Society for Gerontology (ISG'08)*. Pisa (Italy).

Wada, K., Shibata, T., Mushi, T., & Kimura, S. (2008). Robot therapy for elders affected by dementia: Using personal robots for pleasure and relaxation. *IEEE Engineering in Medicine and Biology Magazine*, 53-60.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford; New York: Oxford University Press.

Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36-45.

WHO. (2010). *Health topics: Ageing*, from http://www.who.int/topics/ageing/en/

Winner, L. (1980). Do artifacts have politics? *Daedalus, 109*, 121-136.

Yeung, K. (2008). Towards an understanding of design-based instruments. In R. Brownsword & K. Yeung (Eds.), *Regulating Technologies: Legal futures, regulatory frames and technological fixes* (pp. 79-109). Oxford: Hart.